BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Express Mail Label No.: EV287336791US
Date of Deposit: April 23, 2004

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS:	Jonathan M. Rothberg, et al.		
ASSIGNEE:	CuraGen Corporation		
SERIAL NUMBER:	09/814,338	EXAMINER:	Young J. Kim
FILING DATE:	March 21, 2001	ART UNIT:	1637
FOR:	METHOD OF SEQUENCING A NUCLEIC ACID		

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

DECLARATION OF MARCEL MARGULIES, Ph.D. UNDER 37 C.F.R. §1.132

I, MARCEL MARGULIES, declare and state that:

1. I am Vice President of Engineering, at 454® Life Sciences, the exclusive licensee of this application. My previous employment includes Director of New Technology Research at Perkin-Elmer's Instrument Division in Norwalk, CT, and Associate Director of the Hubble Space Telescope project.

2. I earned my B.Sc. in Engineering from the Free University of Brussels, in Belgium, and a Ph.D. in theoretical physics from Columbia University.

3. I have reviewed the instant application and the November 6, 2003 Office Action in this case.

4. Based on information and belief, it is my opinion that the claimed sequencing apparatus and substrate of the instant application are vastly superior to other sequencing systems, including the system reported by Chee et al. in Published Application No. U.S. 2003/0108867 ("Chee et al."), as cited by the Examiner.

The superior performance of the claimed invention can be attributed to the functional features of the apparatus and substrate, which include: 1) compact wafers; 2) attachable optical fibers; 3) flow chamber and fluid means; and 4) specific fiber and well sizes. As a result of these features (and others), the claimed invention provides the first massively parallel, solid-phase, whole-genome sequencing platform that can be scaled for viral, bacterial, and even human genomes.

Functional features of the claimed substrate and apparatus:
compact wafers can be placed in flow chambers for efficient fluid exchange

5. As recited in the claims, the compact wafers of the invention allow *placement into flow chambers*, which utilize an *efficient fluid exchange system* and thereby provide significantly faster sequence analysis. The advantages of the claimed apparatus and substrate are fully disclosed in the instant application, as filed.

6. The instant application teaches that the claimed apparatus and substrate include a compact wafer formed from a bundle of optical fibers, cut and polished to a thickness of 0.5 mm to 5 mm. This teaching is found, *inter alia*, in the originally filed application on p. 36, l. 12-15; and p. 36, l. 30 to p. 37, l. 3.

7. The instant application additionally teaches that the claimed substrate can be used with a flow chamber and fluid means for delivering sequencing reagents and washes to the wafer surface. This teaching is found, *inter alia*, in the originally filed application in Figs. 2 and 3; on p. 4, l. 14-18; p. 30, l. 15-19; p. 33, l. 20 to p. 35, l. 20; and in Exmp. 3 on p. 53, l. 28 to p. 54, l. 15.

8. The instant application further teaches that the underside of the compact wafer in the flow chamber can be optically linked or directly contacted with a fiber optic bundle to allow image capture, for example, through a CCD system. This teaching is found, *inter alia*, in the originally filed application on p. 34, l. 13-18; and Fig. 2.

9. The instant application teaches also that the underside of the compact wafer in the flow chamber can be placed in proximity to conventional optics mechanism, e.g., a high numerical aperture lens system to allow for image capture. This teaching is found, *inter alia*, in the originally filed application on p. 34, l. 19-23.

10. The claimed wafer for use in the flow chamber allows for much faster sequence analysis. The claimed apparatus and substrate thereby yield significantly improved results, which are not obtained with other sequencing systems such as that reported by Chee et al.

11. As cited by the Examiner, Chee et al. does not specify the length of the optic fibers for their sequencing system. Instead, Chee et al. relies on WO 98/50782 (see Chee et al., ¶ [0007]), which reports the use of optic fibers that are several meters long (Ex. 1). These long and bulky fibers cannot readily fit into a flow chamber. Instead, Chee et al. reports methods of inverting the long optic fiber and sequentially dipping the tip into individual cups filled with solutions of single nucleotides (Chee et al., ¶¶ [0192] – [0195], *inter alia*). This awkward dipping process is necessitated by the long, bulky fiber employed in Chee et al.

12. The dipping method of Chee et al. is predicted to be completely or partly inoperable. First, the continuous plunging of an optic fiber tip into nucleotide solutions would tend to dislodge any beads from the wells. Second, for DNA attached directly to wells, dipping would be ineffective in delivering the nucleotide solution to the wells due to the counteraction of air pressure. This phenomenon is generally observed for inverted cups or glasses placed into reservoirs of water, and provides the basis for the oceanographic apparatus known as the diving bell. An illustration provided as an aid in understanding is included as Ex. 2 (adapted from http://home.earthlink.net/~dmocarski/chapters/chapter7 /main.htm).

13. Even assuming, *arguendo*, that the dipping method of Chee et al. were marginally effective in delivering reagents and preserving samples for sequencing, the method would be extremely slow and inefficient compared to the claimed invention. The long fiber optic tip of Chee et al. would need to be inverted and dipped in and out of at least four cups (e.g., A, T, C, G), or perhaps more than eight cups (e.g., A, first wash, T, second wash, C, third wash, G, fourth wash) to determine only one nucleotide of sequence. In contrast, compact wafer of the invention can be placed into a flow chamber to allow for rapid and efficient delivery of sequencing reagents and washes to the compact wafer. An illustration provided as an aid in understanding is included as Ex. 3.

14. Because the long optic fibers of Chee et al. cannot readily fit into a flow chamber, the sequencing reactions are performed using an unwieldy dipping process. This results in significant delays and increased sample losses. By comparison, the claimed wafer is fitted into a flow chamber to allow streamlined processing and sequence analysis. This is a *significant functional advantage* over the system reported in Chee et al.

Functional features of the claimed substrate and apparatus:
compact wafers have well sizes to maximize signal capture and minimize sample loss

15. As recited in the claims, the compact wafers of the invention include *optimally sized fibers and wells* that allow *maximal signal capture and minimal sample loss* and thereby provide significantly improved sequence analysis. The advantages of the claimed apparatus and substrate are fully disclosed in the instant application, as filed.

16. The instant application teaches that the claimed apparatus and substrate include a compact wafer that includes optic fibers with a diameter of 3 μm to 100 μm. The application teaches that this diameter is important to ensure that each light signal can be captured as a single pixel. This teaching is found, *inter alia*, in the originally filed application on p. 36, l. 15 and 25-29.

17. The instant application teaches that the claimed apparatus and substrate include a compact wafer that includes wells with a depth of one-half to three times the diameter of the optic fibers. The application expressly recognizes the problem of bead/sample loss during the sequencing reaction. This teaching is found, *inter alia*, in the originally filed application on p. 37, l. 6-9; p. 39, l. 22-24; and Fig. 4.

18. The claimed fiber diameter and well depth of the compact wafer allow for much more effective sequence analysis. Accordingly, the claimed apparatus and substrate yield significantly improved results, which are not obtained with other sequencing systems, including the system reported by Chee et al.

19. As cited by the Examiner, Chee et al. ¶ [0105] apparently reports the use of long optic fibers with diameters ranging from approximately 0.17 μm to 0.03 μm. These diameters can be calculated from "high density" arrays indicated by Chee et al. ¶ [0105], i.e., arrays containing 40,000 fibers/mm^2 to 1,000,000 fibers/mm^2 (Office Action, quoting Chee et al. on pages 9-10). Chee et al. appears to be silent as to the specific well depths employed with the optic fibers.

20. With diameters of approximately 0.17 μm to 0.03 μm, many of the optic fibers employed by Chee et al. would produce sequencing systems that are completely or partly inoperable. Optic fibers having a such small diameters would require bead and well sizes less than 0.17 μm to 0.03 μm in diameter. Such systems would be predicted to have a myriad of problems, including difficulties in distinguishing light signals from each fiber and in depositing the beads in the wells. By comparison to Chee et al. ¶ [0105], the compact wafer of the invention employs optic fibers 3 μm to 100 μm in diameter to provide for maximal sample density while still allowing accurate signal detection and efficient bead delivery.

21. Chee et al. do not appear to specify well depths for use with the optic fibers, and evidently fail to recognize the importance of well depth in preventing sample loss.

In fact, Fig. 1 in Chee et al. shows beads and samples jutting out from their wells. This configuration would likely lead to significant sample loss during the "invert and dip" process reported by Chee et al. In contrast, the compact wafer of the invention employs well depths of one-half to three times the diameter of the fiber, which are important in minimizing sample loss during preparation and analysis. The optimally sized fibers and wells (¶¶ 16 and 17, above) therefore represent *significant functional advantages* over the system reported in Chee et al.

Superior function of the claimed substrate and apparatus:
massively parallel analysis of viral and human genomic sequences

22. As a result of these highly advantageous, functional features (¶¶ 14, 20, and 21, above), and other important aspects, the substrate and apparatus claimed in the instant application are the first to allow rapid massively parallel sequencing for whole genomes.

23. Traditional methods for genome sequencing have been slow, expensive, laborious, and industrial-scale, since they involve individually preparing and sequencing DNA fragments of the genome. The Human Genome Project, for example, required approximately 12 years, $2.7 billion dollars, and 60 million samples to complete.

24. In contrast, the substrate and apparatus claimed in the instant application provide a massively parallel, scalable platform that dramatically reduces the time, cost, sample preparation, and space required for genome sequencing. Instead of individually preparing and sequencing each sample, the claimed substrate and apparatus allow parallel sequencing of thousands (or hundreds of thousands) of samples.

25. Recently, the claimed substrate and apparatus were used to sequence the entire adenovirus genome (approximately 30,000 base pairs) contained on an expression vector in less than one day (see NY Times article, Ex. 4). The entire sequencing

process from sample preparation to data analysis was accomplished in less than one day, and provided over 99% genome coverage. The resulting adenovirus sequence was published in GenBank under Accession Nos. AY370909, AY370910, and AY370911 (Ex. 5).

26. In further experiments, the apparatus of the instant application was used to sequence a segment human chromosome 12 (approximately 170,000 base pairs) contained on an artificial chromosome vector (Ex. 6). With the apparatus, a one-day sequencing run produced sufficient shotgun sequence coverage of the chromosome 12 clone (Ex. 6, p. 6). A single sequencing run obtained 85% genome coverage and 98% consensus accuracy (Ex. 6, p. 3). These results were presented at the15th Annual Genome Sequencing and Analysis Conference, held on September 21-24, 2003 (Ex. 6, p. 1).

27. To generate this sequence information described in ¶¶ 25 and 26 (above), preferred commercial embodiments of the claimed substrate and apparatus were fabricated. In these preferred embodiments, the claimed substrates (termed "PicoTiter Plates") were formed from cavitated fiber optic wafers formed from a fused bundle of a plurality of individual optical fibers as taught and claimed by the instant application.

28. Specifically, PicoTiter Plates were made acid etching the top surface of fiber optic wafers to form wells with diameters between 39 and 44 μm, as currently claimed. The fiber optic wafer exhibited a thickness of about 2.0 mm, also as currently claimed. In addition, the wells on PicoTiter Plates were fabricated with depths ranging from 26 to 76 μm (i.e., from between one half the diameter of an individual optical fiber and three times the diameter of an individual optical fiber, as recited in the pending claims). Finally, the wells were loaded with nucleic acid template and beads with pyrophosphate sequencing reagents attached thereto, as recited in the pending claims. Sequencing by synthesis was then performed as

described in the specification, and using the claimed apparatus to flow sequencing reagents over the PicoTiter Plate.

29. The substrate and apparatus claimed in the instant application therefore fulfill a long-felt but unmet need for rapid, whole-genome analysis of viral and bacterial pathogens (e.g., ¶ 25, above). Such analysis is critical for biodefense, drug discovery, and the identification of emerging pathogens. More than this, the claimed apparatus solves the long-standing problems with analysis of large genomes, such as human genomes (e.g., ¶ 26, above). Solutions for large-genome sequencing are vital for drug development, early diagnosis, and faster clinical interventions.

30. For these reasons, in my opinion, the claimed substrate and apparatus represent a significant advancement in the field as the first massively parallel, solid-phase, whole-genome sequencing platform that can be scaled for the smallest to the largest genomes.

Conclusion

31. Therefore, based on information and belief, and all of the foregoing, it is my opinion the claimed sequencing apparatus and substrate substantially outperform the sequencing platforms used by Chee et al. and others. This is due to the functionally superior features of the claimed invention, which include compact fiber optic wafers, detachable fiber optic bundles, flow chambers and fluid means, and specifically sized fibers and wells. All of these features, and the other aspects of the invention, work together to achieve significantly faster results compared to other sequencing systems.

32. I declare that all statements made herein of our own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18

U.S.C. § 1001 and that willful false statements may jeopardize the validity of this application and any patent issuing therefrom.

Dated: 4/20/04

Signed: _____
MARCEL MARGULIES

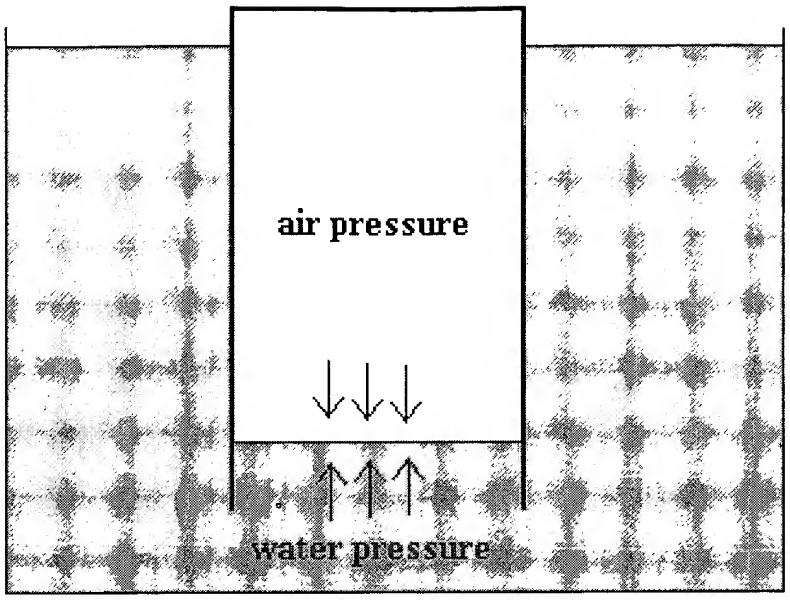
A. The Format Choices

The individually clad, optical fiber strand

A typical optical fiber strand is illustrated by Figs. 1 and 2A and 2B. As seen therein, an individual optical fiber strand 10 is comprised of a single optical fiber 12 having a rod-like shaft 14 and two fiber ends 16, 18, each of which provides a substantially planar end surface. The intended distal surface 20 at the fiber end 16 is illustrated by Fig. 2A, while the intended proximal surface 22 at the fiber end 18 is illustrated within Fig. 2B. It will be recognized and appreciated that the terms "proximal" and "distal" are relative and interchangeable until the strand is ultimately positioned in an apparatus. The optical fiber 12 is composed typically of glass or plastic; and is a flexible rod able to convey light energy introduced at either of its ends 16 and 18. Such optical fibers 12 are conventionally known and commercially available. Alternatively, the user may himself prepare individual optical fibers in accordance with the practices and techniques reported in the scientific and industrial literature. Accordingly, the optical fiber 12 is deemed to be conventionally known and available as such.

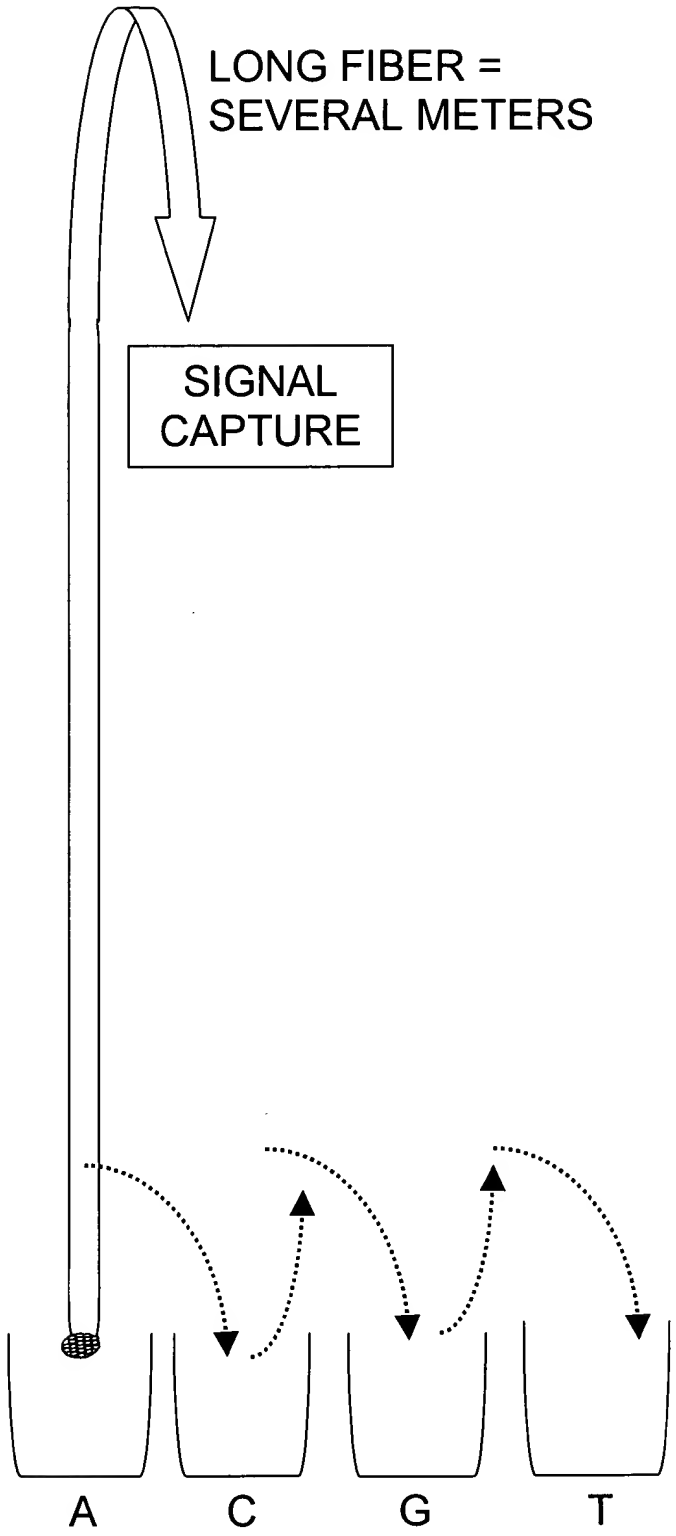
It will be appreciated that Figs. 1-2 are illustrations in which the features have been purposely magnified and exaggerated beyond their normal scale in order to provide both clarity and extreme detail. Typically, the conventional optical fiber has a cross section diameter of 5-500 micrometers; and is routinely employed in lengths ranging between meters (in the laboratory) to kilometers (in field telecommunications). Moreover, although the optical fiber 12 is illustrated via Figs. 1-2 as a cylindrical extended rod having substantially circular proximal and distal end surfaces, there is no requirement or demand that this specific configuration be maintained. To the contrary, the optical fiber may be polygonal or asymmetrically shaped along its length; provided with special patterns and shapes at the proximal and/or distal faces; and need not present an end surface which is substantially planar. Nevertheless, for best efforts, it is presently believed that the substantially cylindrical rod-like optical fiber having planar end surfaces is most desirable.

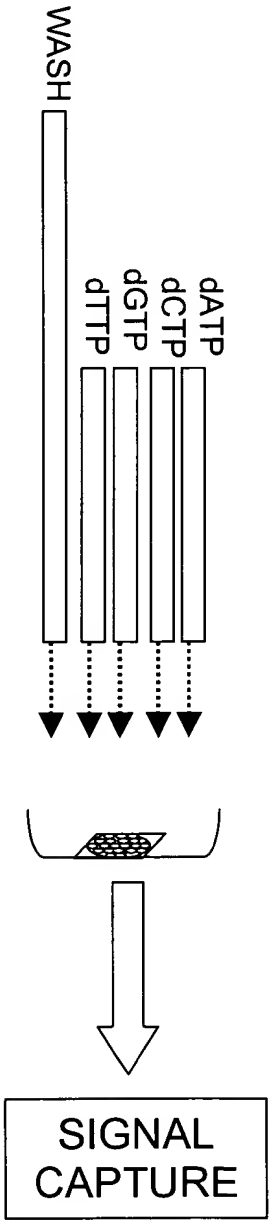
Each optical fiber 12 is desirably, but not necessarily, individually clad axially along its length by cladding 26. This cladding 26 is composed of any material with a lower refractive index than the fiber core and prevents the transmission of light energy photons from the optical fiber 12 to the external environment. The cladding material 26 may thus be composed of a variety of radically different chemical formulations including various glasses, silicones, plastics, platings, and shielding matter of diverse chemical composition


air pressure
water pressure

CHEE et al.

CLAIMED SUBSTRATE





CANNOT USE FLOW CHAMBER – REQUIRES SLOW, WASTEFUL DIPPING PROCESS

CAN USE FLOW CHAMBER – ALLOWS FAST & EFFICIENT PROCESSING

Company Says It Mapped Genes of Virus in One Day

By ANDREW POLLACK

A small company developing a novel method of sequencing genes said yesterday that it had determined the entire genetic code of a virus in a single day.

The company, 454 Life Sciences, said it was the first time that the entire genome of an organism had been sequenced using an unconventional technique and demonstrated the feasibility of its technology, which it said might eventually allow gene sequencing to be done faster and less expensively.

"It's a real threshold moment," said Richard Gibbs, director of the genome-sequencing center at Baylor College of Medicine in Houston and an adviser to the company. "This is going to be big."

Other scientists were more skeptical, saying that sequencing a virus, which has a tiny genome, is trivial and that there are aspects of the

technology that might make it difficult to do more complex organisms.

"I think doing a whole bacterium will be a challenge," said Edward M. Rubin, director of the Joint Genome Institute, a Department of Energy sequencing center in Walnut Creek, Calif. Bacteria are the next level up in complexity after viruses.

454, based in Branford, Conn., is one of several companies racing to improve gene sequencing. Some scientists say it might one day be possible to sequence an entire human genome in a few days for as little as $1,000 so that each person could have his or her genome for use in predicting susceptibility to disease and choosing the most appropriate medicines. The Human Genome Project, which first sequenced a human genome, took years and cost tens of millions of dollars.

Still, the thousand-dollar genome is years away. The human genome, which consists of more than three billion letters of the genetic code, is about 100,000 times the size of the adenovirus that 454 sequenced, which has a genome of about 33,000 letters.

454 is majority owned by CuraGen, a genomics company. The name 454 was the code name by which the project was referred to at CuraGen and the numbers have no special meaning, Richard F. Begley, chief executive of 454, said. Some of the company's technology is licensed from Pyrosequencing A.B., a Swedish company. Mr. Begley said the company hoped to start offering a sequencing service by the end of the year and to begin selling machines about six months after that.

Sequencing is usually done in big sequencing centers with dozens of machines and dozens of people. Dr. Begley said 454's technology would pave the way for the "personalized genome center" in which a single scientist could do sequencing on one

machine in a small room.

The company's technique does more sequencing in parallel than the conventional technique, he said. Another significant time-saving development, he said, comes in preparing the samples for sequencing. The 454 technique, he said, requires one sample preparation for each organism compared with dozens and even thousands for the conventional technique.

Scientists differed on whether sequencing a virus in a day was truly fast. The Department of Energy once did 30 bacteria in 30 days. The virus that causes SARS was sequenced in six days, but it might be misleading to compare that effort with 454's.

RELATED ARTICLES

- A Better Mousetrap Is Built, Detecting Deadly Bacteria (July 11, 2003) $
- Specter Asks Bush to Permit More Embryonic Cell Lines (April 23, 2003) $
- PUBLIC LIVES; A Lab of Dying Mice, and a Virus in His Recipe Box (July 19, 2002) $
- TRACES OF TERROR: THE SCIENCE; SCIENTISTS CREATE A LIVE POLIO VIRUS (July 12, 2002) $
- THE YEAR IN IDEAS; A TO Z.; Attaching Good Genes to Bad Viruses (December 9, 2001) $

Find more results for Viruses and Genetics and Heredity .

TOP BUSINESS ARTICLES

- Schering-Plough Details Its Steps for a Turnaround
- Ouster at Freddie Mac Discussed
- Stocks Surge After Intel Lifts Guidance
- Utilities Divided Over Managing Power Flow
- Abbott Spins Off Hospital Products Unit

☐ **1:** AH013174. Expression vector...[gi:34014916] Links

```
LOCUS       AY370909S1              2617 bp    DNA     linear   SYN 21-AUG-2003
DEFINITION  Expression vector pAdEasy-1, contig 1.
ACCESSION   AY370909
VERSION     AY370909.1  GI:34014917
KEYWORDS    .
SEGMENT     1 of 3
SOURCE      Expression vector pAdEasy-1
  ORGANISM  Expression vector pAdEasy-1
            artificial sequences; vectors.
REFERENCE   1  (bases 1 to 2617)
  AUTHORS   Sarkis,G., Costa,G., Leamon,J., Maithreyan,S., Berka,J., Du,L.,
            Fierro,J., McDade,K., Puc,B., Roth,G.T., Gomes,X., Altman,W.,
            Charumilind,J., Chen,Y.-J., Chen,Z., de Winter,A., Dewell,S.,
            Drake,J., Forte,R., He,W., Helgesen,S., Jannotti,M.L., Jarvie,T.,
            Jirage,K., Kelch,K., Kim,J.-B., Kukanski,K., Lanza,J., Lee,W.,
            Lefkowitz,S., Lu,H., Makhijani,V., Margulies,M., Nobile,J.,
            Norton,W., Reifler,M., Rodgers,G., Ronan,M., Simpson,J.,
            Tartaro,K., Verma,S., Zimmerman,Z., Dacey,P., Begley,R. and
            Lohman,K.
  TITLE     Sequence Analysis of the pAdEasy-1 Recombinant Adenoviral Construct
            Using the 454 Life Sciences Sequencing-by-Synthesis Method
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 2617)
  AUTHORS   Lohman,K.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-AUG-2003) 454 Life Sciences, 20 Commercial Street,
            Branford, CT 06405, USA
FEATURES             Location/Qualifiers
     source          1..2617
                     /organism="Expression vector pAdEasy-1"
                     /mol_type="other DNA"
                     /db_xref="taxon:243021"
                     /clone="Stratagene catalog number 240005"
                     /note="contig 1; differs from pAdEasy-1 sequence from
                     Stratagene; sequenced by new method"
BASE COUNT      576 a    748 c    711 g    582 t
ORIGIN
        1 aattaacatg catggatcct acgtctcgac cgatgccctt gagagccttc aacccagtca
       61 gctccttccg gtgggcggcg gggcatgact atcgtcgccg cacttatgac tgtcttcttt
      121 atcatgcaac tcgtaggaca ggtgccggca gcgctctggg tcattttcgg cgaggaccgc
      181 tttcgctgga gcgcgacgat gatcggcctg tcgcttgcgg tattcggaat cttgcacgcc
      241 ctcgctcaag ccttcgtcac tggtcccgcc accaaacgtt tcggcgagaa gcaggccatt
      301 atcgccggca tggcggccga cgcgctgggc tacgtcttgc tggcgttcgc gacgcgaggc
      361 tggatggcct tccccattat gattcttctc gcttccggcg gcatcgggat gcccgcgttg
      421 caggccatgc tgtccaggca ggtagatgac gaccatcagg gacagcttca aggatcgctc
      481 gcggctctta ccagcctaac ttcgatcatt gttggaccgc tgatcgtcac ggcgatttat
      541 gccgcctcgg cgagcacatg gaacgggttg gcatggattg taggcgccgc cctatacctt
```

```
      601 gtctgcctcc ccgcgttgcg tcgcggtgca tggagccggg ccacctcgac ctgaatggaa
      661 gccggcggca cctcgctaac ggattcacca ctccaagaat tggagccaat caattcttgc
      721 ggagaactgt gaatgcgcaa accaaccctt ggcagaacat atccatcgcg tccgccatct
      781 ccagcagccg cacgcggcgc atctcgggca gcgttgggtc ctggccacgg gtgcgcatga
      841 tcgtgctcct gtcgttgagg acccggctag gctggcgggg ttgccttact ggttagcaga
      901 atgaatcacc gatacgcgag cgaacgtgaa gcgactgctg ctgcaaaacg tctgcgacct
      961 gagcaacaac atgaatggtc ttcggtttcc gttgtttcgt aaagtctgga aacgcggaag
     1021 tcagcgccct gcaccattat gttccggatc tgcatcgcag gatgctgctg gctaccctgt
     1081 ggaacaccta catctgtatt aacgaagcgc tggcattgac cctgagtgat ttttcttctg
     1141 gtcccgccgc atccataccg ccagttgttt accctcacaa cgttccagta accgggcatg
     1201 ttcatcatca gtaacccgtc atcgtgagca tcctctctcg tttcatcggt atcattaccc
     1261 ccatgaacaa gaaatccccc ttacacggag gcatcagtga ccaaacaagg aaaaaaccag
     1321 cccttaacat ggcccgcttt atcagaagcc agacattaac gcttctggag aaactcaacg
     1381 agctggacgc ggatgaacag gcagacatct gtgaatcgct tcacgaccac gctgatgagc
     1441 tttaccgcag ctgcctcgcg cgtttcggtg atgacggtga aaacctctga cacatgcagc
     1501 tcccggagac ggtcacagct tgtctgtaag cggatgccgg gagcagacaa gcccgtcagg
     1561 gcgcgtcagc gggtgttggc gggtgtcggg gcgcagccat gacccagtca cgtagcgata
     1621 gcggcagtgt atactggctt aactatgcgg catcagagca gattgtactg agagtgcacc
     1681 atatgcggtg tgaaataccg cacagatgcg taaggaagaa aataaccgca tcaggcgctc
     1741 ttccgcttcc tcgctcactg actcgctgcg ctcggtcgtt cggctgcggc gagcggtatc
     1801 agctcactca aaggcggtaa tacggttatc cacagaatca ggggataacg caggaaagaa
     1861 catgtgagca aaaggccaag caaaaggcca aggaaccagt aaaaaggccg cgttgctggc
     1921 gtttttccat aggctccgcc cccctgacga gcatcacaaa atcgacgctc aagtcagagg
     1981 tggcgaaacc cgacaggact ataaagatac caggcgtttc ccctggaagc tccctcgtgc
     2041 gctctcctgt taccgaccct gccgcttacc ggatacctgt ccgcctttct tcccttcggg
     2101 aagcgtggcg ctttctcata agctcacgct gtaggtatct cagttcggtg taggtcgttc
     2161 gctccaagct gggctgtgtg cacgaacccc ccgttcagcc cgaccgctgc gccttatccg
     2221 gtaactatcg tcttgagtcc aacccggtaa gacacgactt atcgccactg gcagcagcca
     2281 ctggtaacag gattagcaga gcgaggtatg taggcggtgc tacagagttc ttgaagtggt
     2341 ggcctaacta cggctacact agaaggacag tatttggtat ctgcgctctg ctgaagccag
     2401 ttaccttcgg aaaagagttg gtagctcttg atccggcaaa caaaccaccg ctggtagcgg
     2461 tggtttttg tttgcaagca gcagattacg cgcagaaaa aaggaatctc aagaagattc
     2521 ctttgtattc ttttcttacg gggtgctgac gctcagtgga acgaaaactc acgttaaggg
     2581 attttggtca tgagattatc aaaaaggatc ttcacct
//
LOCUS       AY370909S2              1062 bp    DNA     linear   SYN 21-AUG-2003
DEFINITION  Expression vector pAdEasy-1, contig 2.
ACCESSION   AY370910
VERSION     AY370910.1  GI:34014918
KEYWORDS    .
SEGMENT     2 of 3
SOURCE      Expression vector pAdEasy-1
  ORGANISM  Expression vector pAdEasy-1
            artificial sequences; vectors.
REFERENCE   1  (bases 1 to 1062)
  AUTHORS   Sarkis,G., Costa,G., Leamon,J., Maithreyan,S., Berka,J., Du,L.,
            Fierro,J., McDade,K., Puc,B., Roth,G.T., Gomes,X., Altman,W.,
            Charumilind,J., Chen,Y.-J., Chen,Z., de Winter,A., Dewell,S.,
            Drake,J., Forte,R., He,W., Helgesen,S., Jannotti,M.L., Jarvie,T.,
            Jirage,K., Kelch,K., Kim,J.-B., Kukanski,K., Lanza,J., Lee,W.,
            Lefkowitz,S., Lu,H., Makhijani,V., Margulies,M., Nobile,J.,
            Norton,W., Reifler,M., Rodgers,G., Ronan,M., Simpson,J.,
            Tartaro,K., Verma,S., Zimmerman,Z., Dacey,P., Begley,R. and
            Lohman,K.
  TITLE     Sequence Analysis of the pAdEasy-1 Recombinant Adenoviral Construct
            Using the 454 Life Sciences Sequencing-by-Synthesis Method
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1062)
  AUTHORS   Lohman,K.
  TITLE     Direct Submission
```

```
3121 atccttgcaa gtctagcgcc tgctgccatg cgcgggcggc aagcgcgcgc tcgtatgggt
3181 tgaggtgggg gaccccatgg catggggtgg gtgagcgcgg aggcgtacat gccgcaaatg
3241 tcgtaaacgt aggaggggct ctctgagtat tccaagatat gtagggtagc atcttccacc
3301 gcggatgctg gcgcgcacgt aatcgtatag ttcgtgcgag ggagcgagga ggtcgggacc
3361 gaggttgcta cgggcgggct gctctgctcg gaagactatc tgcctgaaga tggcatgtga
3421 gttggatgat atggttggac gctggaagac gttgaagctg gcgtctgtga gacctaccgc
3481 gtcacgcacg aaggaggcgt aggagtcgcg cagcttgttg accagctcgg cggtgacctg
3541 cacgtctagg gcgcagtagt ccagggtttc cttgatgatg tcatacttat cctgcttccc
3601 tttttttcca cagctcgcgg ttgaggacaa actcttcgcg gtctttccag tactcttgga
3661 tcggaaaccc gtcggcctcc gaacggtaag agcctagcat gtagaactgg ttgacggcct
3721 ggtaggcgca gcatcccttt tctacgggta gcgcgtatgc ctgcgcggcc ttccggagcg
3781 aggtgtgggt gagcgcaaag gtgtccctga ccatgacttt gaggtactgg tatttgaagt
3841 cagtgtcgtc gcatccgccc tgcctcccag aagcaaaagt ccgtgcgctt tttggaacgc
3901 ggatttggca gggcggaagg tgacatcgtt gaagagtatt ctttcccgcg cgaggcataa
3961 agttgcgtgt gatgcggaag ggtcccggca cctcggaacg gttgttaatt acctgggcgg
4021 cgagcacgat ctcgtcaaag ccgttgatgt tgtggcccac aatgtaaagt tccaagaagc
4081 gcgggatgcc cttgatggaa ggcaattttt aagtttcctc gtaggtgagc tcttcagggg
4141 agctgagccc gtgctctgaa agggcccagt ctgcaagatg agggttggaa gcgacgaatg
4201 agctccacag gtcacgggcc attagcattt gcaggtggtc gcgaaaggta cctaaactgg
4261 cgacctatgg cctatttttt cttggggtgg atgcagtaga aggtaagcgg gtcttgttcc
4321 cagcggtccc atccaaggtt cgcggctagg tctcgcgcgg cagtcactag aggctcatct
4381 ccgccgaact tcatgaccag catgaagggc acgagctgct tcccaaaggc ccccatccaa
4441 gtataggtct ctacatcgta ggtgacaaag agacgctcgg tgcgaggatg cgagccgatc
4501 gggaagaact ggatctcccg ccaccaattg gaggagtggc tattgatgtg gtgaaagtaa
4561 gaagtccctg cgacgggccg aacactcgtg cttggctttt gtaaaaacgt gcgcagtact
4621 ggcagcggtg cacgggctgt acatcctgca cgaggttgac ctgacgaccg cgcacaagga
4681 agcagagtgg gaatttgagc ccctccgcct ggcgggtttg gctggtggtc ttctacttcg
4741 gctgcttgtc cttgaccgtc tggctgctcg aggggagtta cggtggatcg gaccaccacg
4801 ccgcgcgagc ccaaagtcca gatgtccgcg cgcggcggtc ggagcttgat gacaacatcg
4861 cgcagatggg agctgtccat ggtctggagc tcccgcggcg tcaggtcagg cgggagctcc
4921 tgcaggttta cctcgcatag acgggtcagg gcggcgggct agatccaggt gatacctaat
4981 ttccaggggc tggttggtgg cggcgtcgat ggcttgcaag aggccgcatc cccgcggcgc
5041 gactacggta ccgcgcggcg ggcggtgggc cggcgggggt gtccttggat gatgcatcta
5101 aaagcggtga cgcgggcgag cccccggagg gtaggggggg ctccggaccc gccgggagga
5161 ggggggcaggg gcacgtcggc gccgcgcgcg ggcaggagct ggtgctgcgc gcgtaggttg
5221 ctggcgaacg cgacgacgcg gcggttgatc tcctgaatct ggcgcctctg cgtgaagacg
5281 acgggcccgg tgagcttgaa aacctgaaag agagttcgac agaatcaatt tcggtgtcgt
5341 tgacggcggc ctggcgcaaa atctcctgca cgtctcctga gttgtcttga taggcgatct
5401 cggccatgaa ctgctcgatc tcttcctcct ggagatctcc gcgtccggct cgctccacgg
5461 tggcggcgag gtcgttggaa atgcgggcca tgagctgcga gaaggcgttg aggcctccct
5521 cgttccagac gcggctgtag accaccgccc ccttccggca tcgcgggcgc gcatgaccac
5581 ctgcgcgaga ttgagctcca cgtgccgggc gaagacggcg tagtttcgca ggcgctgaaa
5641 gaggtagttg agggtggtgg cggtgtgttc tgccacgaag aagtacataa cccagcgtcg
5701 caacgtggat tcgttgatat cccccaaggc ctcaaggcgc tccatggcct cgtagaagtc
5761 cacggcgaag ttgaaaaact gggagttgcg cgccgacacg gttaactcct cctccagaag
5821 acggatgagc tcggcgacag tgtcgcgcac ctcgcgctca aaggctacag gggcctcttc
5881 ttcttcttca atctcctctt ccataagggc ctcccttct tcttcttctg gcggcggtgg
5941 gggaggggggg acacggcggc gacgacggcg caccgggagg cggtcgacaa agcgctcgat
6001 catctccccg ccggcgacgg cgcatggtct cggtgacggc gcggccgttc tcggcggggg
6061 cgcagttgga agacgccgcc cgtcatgtcc cggttatggg ttgggcgggg ggctgccatg
6121 cggcagggat acggcgctaa cgatgcatct caacaattgt tgtgtaggta ctccgccgcc
6181 gagggacctg agcgagtccg catcgaccgg atcggaaaac ctctcgagaa aggcgtctaa
6241 ccagtcacag tcgcaaggta ggctgagcac cgtggcgggc ggcaggcggg cggcggtcgg
6301 ggttgtttct ggcggaggtg ctgctgatga tgtaattaaa gtaggcggtc ttgagacggc
6361 ggatggtcga cagaagcacc atgtccttgg gtccggcctg ctgaatgcgc aggcggtcgg
6421 ccatgcccca ggcttcgttt tgacatcggc gcaggtcttt gtagtagtct tgcatgagcc
6481 tttctaccgg cacttcttct tctccttcct cttgtcctgc atctcttgca tctatcgctg
6541 cggcggcggc ggagtttggc cgtaggtggc gccctcttcc tcccatgcgt gtgaccccga
6601 agcccctcat cggctgaagc agggctaggt cggcgacaac gcgctcggct aatatggcct
6661 gctgcacctg cgtgagggta gactggaagt catccatgtc cacaaagcgg tggtatgcgc
```

```
24721 gtagcctgat tcgggagttt acccagccgc cccctgctag ttgagcggga caggggaccc
24781 tgtgttctca ctgtgatttg caactgtcct aaccctggat tacatcaaga tcctctagtt
24841 aattcagtac ccggggactt acttacctta acccctttaac taaataaaaa aataaataaa
24901 agcatcactt acttaaaatc agttagcaaa tttctgtcca gtttattcag cagcacctcc
24961 cttgccctcc ctcccagctc tggtattgca gcttcctcct ggctgcaaac tttctccaca
25021 atactaaata ggaatgtcag tttcctcctg ttcctgtcca tccgcaccca ctatcttcat
25081 gttgttgcag atgaagcgcg caagaccgtc tgaagatacc ttccaacccc gtgtatccat
25141 atgacacgga aaccggatcc tccaactgtt gccttttctt actcctccct ttgtatcccc
25201 caatgggttt caagagagtc cccctggggt actctctttg cgcctatccg aacctctagt
25261 tacctccaat ggcatgcttg cgctcaaaat gggcaacggc ctctctctgg acgaggccgg
25321 caaccttacc tcccaaaatg taaccactgt gagcccacct ctacaaaaaa ccaaagtcaa
25381 acaataaacc tggaaatatc tgcacccctc acagttacct cagaagccct aactgtggct
25441 gccgccgcac ctctaatggt cgcgggcaac acactcacca tgcaatcaca ggccccgcct
25501 aaccgtgcac gactccaaac ttaagcattg ccacccaagg cacccctcca cagtgtcaga
25561 aggaaagcta gccctgcaaa catcaggccc cctccaccac caccgatagc agtaccctta
25621 ctatcactgc ctccaccccc tcctaactac tgccactggt agcttgggca ttgacttgaa
25681 agagcccatt tatacaacaa aataggaaaa ctaggactaa agtacggggc tcctttgcat
25741 gtaacagacg acctaaacac tttgaccgta gcaactggtc caggtgtgac tattaataat
25801 acttccttgc aaactaaagt tactggagcc ttgggttttg attcacaagg caatatgcaa
25861 cttaatgtag caggaggact aaggattgat tctcaaaaca gacgccttat acttgatgtt
25921 agttatccgt ttgatgctca aaaccaacta aatctaagac taggacaggg cccttctttt
25981 tattaaactc agcccacaac ttggatatta actaacaaca aaggccttta cttgtttaca
26041 gcttcaaaca attccaaaag cttgaggtta acctaagcac tgccaagggg ttgatgtttg
26101 acgctacagc catagccatt aatgcaggag atgggcttga atttggttca cctaatgcaa
26161 ccaaacaaca aactccccta ccaaaacaaa aattaggcca tggcctagaa tttgattcaa
26221 acaaggctat ggttcctaaa ctaggaactg gccttagttt tgacagcaca ggtgccatta
26281 cagtaaggaa acaaaataat gataagctaa ctttgtggac cacaccagct ccatctccta
26341 actgtagaac taaatgcaga gaaagatgct aaactcactt tggtcttaac aaaatgtggc
26401 agtcaaatac ttgctacagt ttctagtttt ggctgttaaa ggcagtttgg ctccaatatc
26461 tggaacagtt acaagtgctc atcttattat aagatttgac agaaatggag tgctactaaa
26521 caattccttc ctggacccag aatattggaa ctttagaaat ggagatctta ctgaaggcac
26581 agcctataca aacgctgttg gatttatgcc taacctatca gcttatacca aaatctcaca
26641 aggtaaaact aagccaaaag taaactattg tcagtctaag tttaacttaa acggagaaca
26701 aaactaaacc tgtaacacta accattacac taaacggtac acaaggaaac aggagacaca
26761 actccaagtg catactctat gtctattttc tatgggactg gtctggccac aactaattaa
26821 tgaaatattt gccacatcct cttacacttt tcatacattg cccaagaata aagaatcgtt
26881 tgttgttatg tttcaacgtg tttatttttc taattgcaga aaatttcaag tcatttttct
26941 attcagtagt atagccccac caccacatag cttatacaga tcaccgtacc ttaatcaaac
27001 tcacagaacc ctagtattca acctgccacc tccctcccaa cacacagagt acacagtcct
27061 ttcctccccg gcctggcctt aaaaagcaat catatcatgg gtaacagaca tattcttagg
27121 tgttatattc cacacggttt ccttgtcgag ccaaacgctc atcagtgata ttaataaact
27181 ccccgggcca gctcacttaa gttcatgtcg ctgtccagct gctgagccac aggctgctgt
27241 ccaacttgcg gttgcttaac gggcggcgaa ggagaagtcc acgcctacat gggggttagg
27301 agtcataatc gtgcatcagg atagggcggt ggtgctgcag cagcgcgcga ataaactgct
27361 gccgccgccg ctccgtcctg caggaataca acatggcagt ggtctcctca gcgatgattc
27421 gcaccgcccg cagcataagg cgccttgtcc tccgggcaca gcagcgcacc ctgatctcac
27481 ttaaatcagc acagtaactg cagcacagca ccacaatatt gttcaaaata cccacagtgc
27541 aaggcgctgt atccaaagct catggcgggg accacagaac ccacgtggcc atcataccac
27601 aagcgcaggt agattaagtg gcgacccctc ataaacacgc tggacataaa cattaccttc
27661 ttttggcatg ttgtaattca ccacctcccg gtaccatata aacctctgat taaacatggc
27721 gccatccacc accatcctaa accagcctag gccaaacctg cccgccggcc ctanactaca
27781 ctgcagggaa ccgggactgg aacaatgaca gtggagagcc caggactcgt aaccatggat
27841 catcatgctc gtcatgatat caatgttggc acaacacagg cacacgtgca tacacttcct
27901 caggattaca agctcctccc gcgttagaac catatcccag ggaacaaccc attcctgaat
27961 cagcgtaaat cccacactgc agggaagacc tcgcacgtaa ctcacgttgt gcattgtcaa
28021 agtgttacat tcgggcagca gcggatgatc ctccagtatg gtagcgcggg tttctgtctc
28081 aaaaggaggt agacgatccc tactgtacgg agtgcgccga gacaaccgag atcgtgttgg
28141 tcgtagtcgt catgccaaat ggaacgccgg acgtagtcat atttcctgaa gcaaaaccaa
28201 ggtgcgggcg tgacaaacag atctgcgtct ccggtctcgc cgcttagatc gctctgtgta
28261 gtagttgtag tatatccact ctctcaaagc atccaggccg ccccctggct tcgggttcta
```

```
  JOURNAL   Submitted (18-AUG-2003) 454 Life Sciences, 20 Commercial Street,
            Branford, CT 06405, USA
FEATURES             Location/Qualifiers
     source          1..1062
                     /organism="Expression vector pAdEasy-1"
                     /mol_type="other DNA"
                     /db_xref="taxon:243021"
                     /clone="Stratagene catalog number 240005"
                     /note="contig 2; differs from pAdEasy-1 sequence from
                     Stratagene; sequenced by new method"
BASE COUNT      282 a    253 c    236 g    290 t      1 others
ORIGIN
        1 aaattaaaat gaatgtttta aatcaatcta aagtatatat gagtaaactt ggtctgacag
       61 ttaccaatgc ttaatcagtg aggcacctat ctcagcgatc tgtcttattt cgttcatcca
      121 tagttgcctg actccccgtc gtgtagataa ctacgatacg ggagggctta ccatctggcc
      181 ccagtgctgc aatgataccg cgagacccac gctcaccggc tccagattta tcagcaataa
      241 accagccagc cggaagggcc gagcgcagaa gtggtcctgc aactttatcc gcctccatcc
      301 agtctattaa ttgttgccgg gaagctagag taagtagttc gccagttaat agtttgcgca
      361 acgttgttgc cattgctgca ggcatcgtgg tgtcacgctc gtcgtttggt atggcttcat
      421 tcagctccgg ttcccaacga tcaaggcgag ttacatgatc ccccatgttg tgcaaaaagc
      481 ggttagctcc ttcggtcctc cgatcgttgt cagaagtaag ttggccgcag tgttatcact
      541 catggttatg gcagcactgc ataattctct tactgtcatg ccatccgtaa gatgcttttc
      601 ttgtgactgg tgagtactca accaagtcat tctgagaata gtgtatgcgg cgaccgagtt
      661 gctcttgccc ggcgtcaaca cgggataata ccgcgccaca tagcaagaaa ctttaaaagt
      721 tagctcatca ttaggaaaac gttcttcggg gcgaaaactc tcaaggatct taccgctgtt
      781 gagatccagt tcgatgtaac ccactcgtgc acccaactga tcttcagcat cttttacttt
      841 caccagcgtt tctgggtgag caaaaacagg aaggcaaaat gccgcaaaaa agggaataag
      901 ggcgacacgg aaatgttgaa tactcatact cttccttttt caatattatt gaagcattta
      961 tcagggttat tgtctcatga gcggatacat atttgaatgt atttagaaaa ataaacaaat
     1021 aggggttccg cgcacatttc cccgaaagtg ccacctgtcn ag
//
LOCUS       AY370909S3             30091 bp    DNA     linear   SYN 21-AUG-2003
DEFINITION  Expression vector pAdEasy-1, contig 3.
ACCESSION   AY370911
VERSION     AY370911.1  GI:34014919
KEYWORDS    .
SEGMENT     3 of 3
SOURCE      Expression vector pAdEasy-1
  ORGANISM  Expression vector pAdEasy-1
            artificial sequences; vectors.
REFERENCE   1  (bases 1 to 30091)
  AUTHORS   Sarkis,G., Costa,G., Leamon,J., Maithreyan,S., Berka,J., Du,L.,
            Fierro,J., McDade,K., Puc,B., Roth,G.T., Gomes,X., Altman,W.,
            Charumilind,J., Chen,Y.-J., Chen,Z., de Winter,A., Dewell,S.,
            Drake,J., Forte,R., He,W., Helgesen,S., Jannotti,M.L., Jarvie,T.,
            Jirage,K., Kelch,K., Kim,J.-B., Kukanski,K., Lanza,J., Lee,W.,
            Lefkowitz,S., Lu,H., Makhijani,V., Margulies,M., Nobile,J.,
            Norton,W., Reifler,M., Rodgers,G., Ronan,M., Simpson,J.,
            Tartaro,K., Verma,S., Zimmerman,Z., Dacey,P., Begley,R. and
            Lohman,K.
  TITLE     Sequence Analysis of the pAdEasy-1 Recombinant Adenoviral Construct
            Using the 454 Life Sciences Sequencing-by-Synthesis Method
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 30091)
  AUTHORS   Lohman,K.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-AUG-2003) 454 Life Sciences, 20 Commercial Street,
            Branford, CT 06405, USA
FEATURES             Location/Qualifiers
     source          1..30091
```

```
                     /organism="Expression vector pAdEasy-1"
                     /mol_type="other DNA"
                     /db_xref="taxon:243021"
                     /clone="Stratagene catalog number 240005"
                     /note="contig 3; differs from pAdEasy-1 sequence from
                     Stratagene; sequenced by new method"
BASE COUNT     6985 a   8742 c   8241 g   6122 t      1 others
ORIGIN
        1 caggtaggaa gaagtagtat aaggtggggt cttatgtagt tttgtatctt gttttgcagc
       61 agccgccgcc gccatgagca ccaactcgtt tgatggaagc attgtgagct catatttgac
      121 aacgcgcatg cccccatggg ccggggtggc gtcagaatgt gatgggctcc agcattgatg
      181 gtcgccccgt cctgcccgca aactctacta ccttgaccta cgagaccgtg tctggaacgc
      241 cgttggagac tgcagcctcc gccgccgctt cagccgctgc agccaccgcc cgcgggattg
      301 tgactgactt tgctttcctg agcccgcttg caagcagtca gcagcttccc gttcatccgc
      361 ccgcgatgac aagttgacgg ctcttttggc acaattggat tctttgaccc gggaacttaa
      421 tgtcgtttcg tcagcagctg ttggatctgc gccagcaggt ttctgccctg aaggcttcct
      481 cccctcccaa tgcggtttaa aacaataaaa taaaaaacca agactctgtt tggtatttgg
      541 atcaagcaag tgtcttgctg tctttattta ggggttttgc gcgcgcggta ggcccgggac
      601 cagcggtctc ggtcgttgag ggtcctgtgt tatttttcca ggacgtggta aaggtgactc
      661 tggatgttca gatacatggg cataagcccg tctctggggt ggaggtagca ccactgcaga
      721 gcttcatgct ggcggggtgg tgttgtagat gatccagtcg tagcaggagc ggctgggcgt
      781 ggtgcctaaa aatgtctttc agtagcaagc tgattgccag gggcaggccc ttggtgtaag
      841 tgtttacaaa gcggttaagc tgggatgggt gcatacgtgg ggatatgaga tgcatcttgg
      901 actgttattt ttaggtttgg ctatgttccc agccatatcc ctccggggat tcatgttgtg
      961 cagaaccacc agcacagtgt atccggtgca cttgggaaat ttgtcatgta gcttagaagg
     1021 aaatgcgtgg aagaacttgg agacgcccct gtgacctcca agattttcca tgcattcgtc
     1081 cataatgatg gcaatgggcc cacgggcggc ggcctgggcg aagatatttc tgggatcact
     1141 aacgtcatag ttgttgttcc aggatgagat cgtcataggc cattttaca aagcgcgggc
     1201 gggagggtgc cagactgcgg tataatggtt ccatccggcc caggggcgta gttaccctca
     1261 cagatttgca tttcccacgc tttgagttca gatgggggga tcatgtctac ctgcggggcg
     1321 atgaagaaaa cggtttccgg ggtaggggag atcagctggg aagaaagcag gttcctgagc
     1381 agctgcgact taccgcagcc ggtgggcccg taaatcacac ctattaccgg cctgcaactg
     1441 gtagttaaga gagctgcagc tgccgtcatc cctgaggcag gggggccact tcgttaagca
     1501 tgtccctgac tcgcatgttt tccctgacca aatccgccag aaggcgctcg ccgcccagcg
     1561 atagcagttc ttgcaaggaa gcaaagtttt tcaacggttt gagaccgtcc gccgtaggca
     1621 tgcttttgag cgtttgacca agcagttcca ggcggtccca cagctcggtc acctgctcta
     1681 cggcatctcg atccagcata tctcctcgtt tcgcgggttg gggcggcttt cgctgtacgg
     1741 cagtagtcgg tgctcgtcca gacgggccag ggtcatgtct ttccacgggc ggcagggtcc
     1801 tcgtcagcgt agtctgggtc acggtgaagg ggtgcgctcc gggctgcgcg ctggccaggg
     1861 tgcgcttgag gctggtcctg ctggtgctga agcgctgccg gtcttcgccc tgcgcgtcgg
     1921 ccaggtagca tttgaccgat ggtgtcatag tccagcccct ccgcggcgtg gcccttggcg
     1981 cgcagcttgc ccttggagga ggcgccgcac gaggggcagt gcagactttt gagggcgtag
     2041 agcttgggcg cgagaaatac cgattccggg gaggtaggca tccgcgccgc caggccccgc
     2101 agacggtctc gcattccacg agccaggtga gctctggccg ttcggggtca aaaaccaggc
     2161 tttcccccat tgctttttga tgcgtttctt acctctggtt tccatgagcc ggtgtccacg
     2221 ctcggtgacg aaaaggctgt ccgtgtcccc gtatacagac ttgagaggcc tgtcctcgag
     2281 cggtgttccg cggtcctcct cgtatagaaa ctcggaccac tctgagacaa aggctcgcgt
     2341 ccaggccagc acgaaggagg ctaaggtggg aggggtaggc ggtcgttgtc cactaggggg
     2401 tccactcgct ccagggtgtg aagacacatg tcgccctctt cggcatcaag gaaggtgatt
     2461 ggtttgtagg tgtaggccac gtgaccgggt gttcctggaa gggggggctaa gtaaaagggg
     2521 gtgggggcgg cgttcgtcct cactctcttc cgcatcgctg tctgcgaggg ccagctgttg
     2581 gggtgagtac tccctctgaa aagcgggcat gacttctgcg ctaagattgt cagtttacca
     2641 aaaacagagg aggatttgat attcacctgg cccgcggtga tgcctttgag ggtggccgca
     2701 tccatctggt caagaaaaga acaatctttt gttgtcaagc ttggtggcaa acgacccgta
     2761 gagggcgttg gacagcaact tggcgatgga gcggcagggt ttggtttttg ttcgcgatcg
     2821 gcgcgctcct tggccgcgat gtttagctgc acgtattcgc gcgcaacgca ccgccattcg
     2881 ggaaagacgg tggtgcgctc gtcgggcacc aggtgcacgc gccaaccgcg gttgtgcagg
     2941 gtgacaaggt caacgctggt ggctacctct ccgcgtaggc gctcgttggt ccagcagagg
     3001 cggccgccct tgcgcgagca gaatggcggg tagggggtct agctgcgtct cgtgccgggg
     3061 ggtctgcgtc cacggtaaag accccgggca gcaggcgcgc gtcgaagtag tctatcttgc
```

```
 6721 ccgtgttgat ggtgtaagtg cagttggcca taacggacca gttaacggtc tggtgacccg
 6781 gctgcgagag ctcggtgtac ctgagacgcg agtaagccct cgagtcaaat acgtagtcgt
 6841 tgcaagtccg caccaggtac tggtatccca ccaaaagtgc ggcggcggct ggcggtagga
 6901 ggggccaggc gtagggtggg ccggggctgc cgggggcgga gatcttccaa cataaggcga
 6961 tgatatccgt agatgtacct ggacatccag gtgatgccgg cggcggtggt ggaggcgcgc
 7021 ggaaagtcgc ggacgcggtt ccagatgttg cgcagcaggc aaaagtagct ccatggtcgg
 7081 gacgctctgg ccggtcaggc gcgcgcaatc gttgacgctc tagcgtgcaa aaggagagcc
 7141 tgtaagcggg cactcttccg tggtctggtg gataaattcg caagggtatc atggcggacg
 7201 accggggttc gagccccgta tccggccgtc cgccgtgatc catgcggtta ccgcccgcgt
 7261 gtcgaaccca ggtgtgcgac gtcagacaag cgggggagtg cttccttttg gcttccttcc
 7321 aggcgcggcg gctgctgcgc ttagcttttt ggccactggc cgcgcgcagc gtaagcggtt
 7381 aggctggaaa gcgaaagcat taagtggctc gctccctgta gccggagggt tattttccaa
 7441 gggttgagtc gcgggacccc cggttcgagt ctcggaccgg ccggactgcg gcgaacgggg
 7501 gtttgcctcc ccgtcatgca agaccccgct tgcaaattcc tccggaaaca gggacgagcc
 7561 cctttttgc cttttcccag atgcatccgg tgctgcggca gatgcgcccc cctccctcag
 7621 cagcggcaag agcaagagca gcggcagaca tgcagggcac cctcccctcc ctcctaccgc
 7681 gtcaggaggg gcgacatccg cggttgacgc ggcagcagat ggtgattacg aacccccgcc
 7741 ggcgccgggc ccggcactac ctggacttgg aggagggcgg agggcctggc gcggctagga
 7801 gcgccctctc ctgagcggcc cacccaaggg atgcagctga agcgtgatac gcgtgaggcg
 7861 tacgtgccgc ggcagaacct gtttcgcgac cgcgagggag aggagcccga ggagatgcgg
 7921 gatcgaaagt tccacgcagg gcgcgagctg cggcatggcc tgaatcgcga gcggttgctg
 7981 cgcgaggagg actttgagcc cgacgcgcga accgggatta gtcccgcgcg cgcacacgtg
 8041 gcggccgccg acctggtaac cgcatacgag cagacggtga accaggagat taactttaca
 8101 aaaagcattt aaacaaccac gtgcgtacgc ttgtggcgcg cgaggaggtg gctataggac
 8161 tgatgcatct gtgggacttt gtaagcgcgc tggagcaaaa cccaaataag caagccgctc
 8221 atggcgcagc tgttccttat agtgcagcac agcagggaca acgaggcatt cagggatgcg
 8281 ctgctaaaca tagtagagcc cgagggccgc tggctgctcg atttgataaa catcctgcag
 8341 agcatagtgg tgcaggagcg cagcttgagc ctggctgaca aggtggccgc catcaactat
 8401 tccatgctta gcctgggcaa gttttacgcc cgcaagatat accatacccc ttacgttccc
 8461 atagacaagg aggtaaagat cgaggggttc tacatgcgca tggcgctgaa tggtgcttac
 8521 cttgagcgac gacctgggcg tttatcgcaa cgagcgcatc cacaaggccg tgagcgtgag
 8581 ccggcggcgc gagctcagcg accgcgagct gatgcacagc ctgcaaaggg ccctggctgg
 8641 cacgggcagc ggcgatagag aggccgagtc ctactttgac gcgggcgctg acctgcgctg
 8701 ggccccaagc cgacgcgccc tggaggcagc tggggccgga cctgggctgg cggtggcacc
 8761 cgcgcgcgct ggcaacgtcg gcggcgtgga ggaatatgac gaggacgatg agtacgagcc
 8821 agaggacggc gagtactaag cggtgatgtt tctgatcaga tgatgcaaga cgcaacggac
 8881 ccggcggtgc gggcggcgct gcagagccag ccgtccggcc ttaactccac ggacgactgg
 8941 cgccaggtca tggaccgcat catgtcgctg actgcgcgca atcctgacgc gttccggcag
 9001 cagccgcagg ccaaccggct ctccgcaatt ctggaagcgg tggtcccggc gcgcgcaaac
 9061 cccacgcacg agaaggtgct ggcgatcgta aacgcgctgg ccgaaaacag ggccatccgg
 9121 cccgacgagg ccggcctggt ctacgacgcg ctgcttcagc gcgtggctcg ttacaacagc
 9181 ggcaacgtgc agaccaacct ggaccggctg gtgggggatg tgcgcgaggc cgtggcgcag
 9241 cgtgagcgcg cgcagcagca gggcaacctg ggctccatgg ttgcactaaa cgccttcctg
 9301 agtacacagc ccgccaacgt gccgcgggga caggaggact acaccaactt tgtgagcgca
 9361 ctgcggctaa tggtgactga gacaccgcaa agtgaggtgt accagtctgg gccagactat
 9421 ttttccagac cagtagacaa ggcctgcaga ccgtaaacct gagccaggct ttcaaaaact
 9481 tgcaggggct ggtggggggt ggcgggctcc cacaggcgac cgcgcgaccg tgtctagctt
 9541 gctgacgccc aactcgcgcc tgttgctgct gctaatagcg cccttcacgg acagtggcag
 9601 cgtgtcccgg gacacatacc taggtcactt gctgacactg taccgcgagg ccataggtca
 9661 ggcgcatgtg gacgagcata ctttccagga gattacaagt gtcagccgcg cgctggggca
 9721 ggaggacacg ggcagcctgg aggcaaccct aaactacctg ctgaccaacc ggcggcagaa
 9781 gatcccctcg ttgcacagtt taaacagcga ggaggagcgc attttgcgct acgtgcagca
 9841 gagcgtgagc cttaacctga tgcgcgacgg ggtaacgccc agcgtggcgc tggacatgac
 9901 cgcgcgcaac atggaaccgg gcatgtatgc ctcaaaccgg ccgtttatca accgcctaat
 9961 ggactacttg catcgcgcgg ccgccgtgaa ccccgagtat ttcaccaatg ccatcttgaa
10021 cccgcactgg ctaccgcccc ctggtttcta caccgggga ttcgaggtgc ccgagggtaa
10081 cgatggattc ctctgggacg acatagacga cagcgtgttt tccccgcaac cgcagaccct
10141 gctagagttg caacagcgcg agcaggcaga ggcggcgctg cgaaaggaaa gacttccgca
10201 ggccaagcag cttgtccgat ctaggcgctg cggccccgcg gtcagatgct agtagcccat
10261 ttccaagctt gatagggtct cttaccagca ctcgcaccac ccgcccgcgc ctgctgggcg
```

```
10321 aggaggagta cctaaacaac tcgctgctgc agccgcagcg cgaaaaaacc tagcctccgg
10381 catttcccaa ccaacgggat agagagccta gtggacaaga tgagtagatg gaagacgtac
10441 gcgcaggagc acagggacgt gccaggcccg ccgcccgccc acccgtcgtc aaaggcacga
10501 ccgtcagcgg ggtctggtgt gggaggacga tgactcggca gacgacagca gcgtcctgga
10561 tttgggaggg agtggcaacc cgtttgcgca ccttcgcccc aggcctgggg aggaatagtt
10621 ttaaaaaaaa aaaaatagat gcaaaataaa aaactacacc aaggccatgg caccgagcgt
10681 tggttttctt gtattccccc tagtatgcgg cgcgcggcga tgtatgagga aggtcctcct
10741 ccctcctacg agagtgtggt gagcgcggcg ccagtggcgg cggcggctgg gttctccctt
10801 cgatgctccc ctggacccgc cgtttgtgcc tccgcggtac ctgcggccta ccgggggggag
10861 aaacagcatc cgttactctg agttggcacc cctattcgac accacccgtg tgtacctggt
10921 ggacaacaag tcaacggatg tggcatccct gaactaccag aacgaccaca gcaactttct
10981 gaccacggtc attcaaaaca atgactacag cccgggggag gcaagcacac agaccatcaa
11041 tcttgacgac cggtcgcaca tggggcggcg acctgaaaac catcctgcat accaacatgc
11101 caaatgtgaa cgagttcatg tttaccaata agtttaaggc ggcgggtgat ggtgtcgcgc
11161 ttgcctacta aggacaatca ggtggagctg aaatacgagt gggtggagtt cacgctgccc
11221 gagggcaact actccgagac catgaccata gaccttatga acaacgcgat cgtggagcac
11281 tacttgaaag tgggcagaca gaacggggtt gctggaaagc gacatcgggg taaaggtttg
11341 acacccgcaa cttcagactg gggtttgacc ccgtcactgg tcttgtcatg cctggggtat
11401 atacaaacga agccttccat ccagacatca ttttgctgcc aggatgcggg gtggacttca
11461 cccacagccg cctgagcaac ttgttgggca tccgcaagcg gcaacccttc caggagggct
11521 ttaggatcac ctacgatgat ctggagggtg gtaacattcc cgcactgttg gatgtggacg
11581 cctaccaggc gagcttgaaa gatgacaccg aacagggcgg gggtggcgca ggcggcagca
11641 acagcagtgg cagcggcgcg gaagagaact ccaacgcggc agccgcggca atgcagccgg
11701 tggaggacat gaacgatcat gccattcgcg gcgacacctt tgccacacgg gctgaggaga
11761 agcgcgctga ggccgaagca gcggccgaag ctgccgcccc cgcctgcgca acccgaggtc
11821 gagaagcctc agaagaaacc ggtgatcaaa cccctgacag aggacagcaa gaaacgcagt
11881 tacaacctaa taagcaatga cagcaccttc acccagtacc gcagctggta ccttgcatac
11941 aactacggcg accctcagac cggaatccgc tcatggaccc tgctttgcac tcctgacgta
12001 acctgcggct cggagcaggt ctactggtcg ttgccagaca tgatgcaaga ccccgtgacc
12061 ttccgctcca cgcgccagat cagcaacttt ccggtggtgg gcgccgagct gttgcccgtg
12121 cactccaaga gcttctacaa cgaccaggcc gtctactccc aactcatccg ccagtttacc
12181 tctctgaccc acgtgttcaa tcgctttccc gagaaccaga ttttggcgcg cccgcccagc
12241 ccccaccatc accaccgtca gtgaaaacgt tcctgctctc acagatcacg ggacgctacc
12301 gctgcgcaac agcatcggag gagtccagcg agtgaccatt actgacgcca gacgccgcac
12361 ctgcccctac gtttacaagg ccctgggcat agtctcgccg cgcgtcctat cgagccgcac
12421 ttttgagcaa gcatgtccat ccttatatcg cccagcaata acacaggctg gggcctgcgc
12481 ttcccaagca agatgtttgg cggggccaag aagcgctccg accaacaccc agtgcgcgtg
12541 cgcgggcact accgcgcgcc ctggggcgcg cacaaacgcg gccgcactgg gcgcaccacc
12601 gtcgatgacg ccatcgacgc ggtggtggag gaggcgcgca actacacgcc caccgccgcc
12661 accagtgtcc acagtggacg cggccattca gaccgtggtg cgcggagccc ggcgctatgc
12721 taaaatgaag agacggcgga ggcgcgtagc acgtcgccac cgccgccgac ccggcactgc
12781 cgcccaacgc gcggcggcgg ccctgcttaa ccgcgcacgt cgcaccggcc gacgggcggc
12841 catgcgggcc gctcgaaggc tggccgcggg tattgtcact gtgccccccа ggtccaggcg
12901 acgagcggcc gccgcagcag ccgcggccat tagtgctatg actcagggtc gcaggggcaa
12961 cgtgtattgg gtgcgcgact cggttagcgg cctgcgcgtg cccgtgcgca ccccgccccc
13021 cgccgcaact agattgcaag aaaaaactaa cttagactcg tactgttgta tgtatccagc
13081 ggcggcggcg cgcaacgaag ctatgtccaa gcgcaaaatc aaagaagaga tgctccaggt
13141 catcgcgccg gagatctatg gccccccgaa gaaggaagag caggattaca agccccgaaa
13201 gctaaagcgg gtcaaaaaga aaaagaaaag aatgatgatg atgaacttga cgacgaggtg
13261 gaactgctgc acgctaccgc gcccaggcga cgggtacagt ggaaaggtcg acgcgtaaaa
13321 cgtgttttgc gacccggcac caccgtagtc tttacgcccg gtgagcgctc cacccgccac
13381 ctacaagcgc gtgtatgatg aggtgtacgg cgacgaggac ctgcttgagc acggccaacg
13441 agcgcctcgg ggagtttgcc tacggaaagc ggcataagga catgctggcg ttgccgctgg
13501 acgagggcaa cccaacacct agcctaaagc ccgtaacact gcagcaggtg ctgcccgcgc
13561 ttgcaccgtc cgaagaaaag cgcggcctaa agcgcgagtc tggtgacttg gcacccaccg
13621 tgcagctgat ggtacccaag cgccagcgac tggaagatgt cttggaaaaa atgaccgtgg
13681 aacctgggct ggagcccgag gtccgcgtgc ggccaatcaa gcaggtggcg ccgggactgg
13741 gcgtgcagac cgtggacgtt cagataccca ctaccagtag caccagtatt gccaccgcca
13801 cagagggcat ggagacacaa acgtccccgg ttgcctcagc ggtggcggat gccgcggtgc
13861 aggcggtcgc tgcggccgcg tccaagacct ctacggaggt gcaaacggac ccgtggatgt
```

```
13921 ttcgcgtttc cagcccccg gccgcccgcg ccgttcgagg aagtacggcg ccgccagcgc
13981 gctactgccc gaatatgccc tacatccttc cattgcgcct acccccggct atcgtggcta
14041 cacctaccgc cccagaagac gagcaactac ccgacgccga accaccactg gaacccgccg
14101 ccgccgtcgc cgtcgccagc ccgtgctggc cccgatttcc gtgcgcaggg tggctcgcga
14161 aggaggcagg accctggtgc tgccaacagc gcgctaccac cccagcatcg tttaaaagcc
14221 ggtctttgtg gttcttgcag atatggccct cacctgccgc ctccgtttcc cggtgccggg
14281 attccgagga agaatgcacc gtaggagggg catggccggc cacggcctga cgggcggcat
14341 gcgtcgtgcg caccaccggc ggcggcgcgc gtcgcaccgt cgcatgcgcg gcggtatcct
14401 gcccctcctt attccactga tcgccgcggc gattggcgcc gtgcccggaa ttgcatccgt
14461 ggccttgcag gcgcagagac actgaattaa aaacaagttg catgtggaaa ataacaaaat
14521 aaaaagtact ggactctcac gctcgcttgg tcctgtaact attttgtaga atggaagaca
14581 tcaactttgc gtctctggcc ccgccgacac ggctcgcgcc cgttcatggg aaactggcaa
14641 gatatcggca ccagcaatat gagcggtggc gccttcagct ggggctcgct gtggagcggc
14701 aattaaaaat ttcggttcca ccgttaagaa ctatggcagc aaggcctgga acagcagcac
14761 aggccagatg ctgagggata agttgaaaga agcaaaattt ccaacaaaag gtggtagatg
14821 gcctggcctc tggcattagc ggggtgggtg gacctggcca accaggcagt gcaaaataag
14881 attaacagta agcttgatcc ccgccctccc gtagaggagc ctccaccggc cgtggagaca
14941 gtgtctccag aggggcggtg gcgaaaagcg tccgccgccc cgacagggaa gaaactctgg
15001 tgacgcaaat agacgagcct ccctcgtacg aggaggcact aaagcaaggc ctgcccacca
15061 cccgtcccat cgcgcccatg gctaccggag tgctgggcca gcacacaccc gtaacgctgg
15121 acctgccctc cccccgcccg acacccagca gaaacctgtg ctgccaggcc cgaccgccgt
15181 tgttgtaacc cgtcctagcc gcgcgtccct gcgccgcgcc gccagcggtc cgcgatcgtt
15241 gcggcccgta gccagtggca actggcaaag cacactgaac agcatcgtgg gtgctggggg
15301 tgcaatccct gaagcgccga cgatgcttct gatagctaac gtgtcgtatg tgtgtcatgt
15361 atgcgtccat gtcgccgcca gaggagctgc tgagccgccg cgcgcccgct ttccaagatg
15421 gctacccctt cgatgatgcc gcagtggtct tacatgcaca tctcgggcca ggacgcctcg
15481 gagtacctga gccccgggct ggtgcagttt gcccgcgcca ccgagacgta cttcagcctg
15541 aataacaagt ttaagaaacc ccacggtggc gcctacgcac gacgtgacca cagaccggtc
15601 ccagcgtttg acgctgcggt tcatccctgt ggaccgtgag gatactgcgt actcgtacaa
15661 ggcgcggttc accctagctg tgggtgataa ccgtgtgctg gacatggctt ccacgtactt
15721 tgacatccgc ggcgtgctgg acaggggccc tacttttaag cccttactct ggcactgcct
15781 acaacgccct ggcctcccaa gggtgcccca aatccttgcg aatgggatga agctgctact
15841 gctcttgaaa taaacctaag aagaagagga cgatgacaac gaagacgaag tagacgagca
15901 agctgagcaa gcaaaaacta cacgtatttg ggcaggcgcc ttattctggt ataaatatta
15961 caaaggaggg tattcaaata ggtgtcgaag gtcaaacaac ctaaatatgc cgataaaaca
16021 atttcaacct gaacctcaaa taaggagaat ctcagtggta cgaaacaaga aattaaatca
16081 tgcagctggg aggagtacct aaaaaagcac taccccaatg aaaccatgtt acggttcata
16141 tgcaaaaccc acaaatgaaa atggagggca aggcattctt gtaaagcaac aaaatggaaa
16201 gctagaaagt aaagtggaaa tgaatttttt ctaactaact agtaggcagg ccagccgcag
16261 gcaatggtga ttaacttact acctaagtgg tattgtacag tgaagatgta gatataagaa
16321 accccagaca ctcatatttc ttacatgccc actattaagg aaggtaactc acgagaacta
16381 atgggccaac aatctatgcc caacaggcct aattacattg cttttaggga caatttttatt
16441 ggtctaatgt attacaacag cacgggtaat atgggtgttc tggcgggcca agcatcgcag
16501 ttgaatgctg ttgtagattt gcaagacaag aaacaacaga gctttcatac cagcttttgc
16561 ttgattccat tggtgataga accaggtact tttctatgtg gaatcaggct gttgacagct
16621 atgatccaga tgttagaatt attgaaatac atggaactag aagatgaact taccaaatta
16681 ctgctttcca ctgggaggtg tgattaatac agagactctt accaaggtaa aacctaaaac
16741 aggtcaggaa aataggatgg gaaaagaata gctacagaat tttacttaga taaaataaga
16801 aataagagtt ggaaataatt ttgccatgga aatcaatcta aatgccaacc tgtggagaaa
16861 tttcctgtac tccaacatag cgctgtattt gcccgacaag ctaaagtaca gtccttccaa
16921 cagtaaaaat ttctgataac ccaaacaacc tacgactaca tgaacaagcg agtggtggct
16981 cccgggcctc agtggactgc tacattaacc ttggagcacg ctggtccctt gactatatgg
17041 acaacgtcaa cccatttaac caccaccgca atgctggcct gcgctaccgc tcaatgttgc
17101 tgggcaatgg tcgctatgtg cccttccaca tccaggtgcc tcagaagttc tttgccatta
17161 aaaacctcct tctcctgccg ggctcataca cctacgagtg gaacttcagg aaggatgtta
17221 acatggttct gcagagctcc ctaggaaatg acctaagggt tgacggagcc agcattaagt
17281 ttgatagcat ttgcctttac gccaccttcc ttccccatgg cccacaacac cgcctccacg
17341 cttgaggcca tgcttagaaa cgacaccaac gaccagtcct ttaacgacta tctctccgcc
17401 gccaacatgt ctctacccta tacccgccaa cgctaccaac gtgcccatat ccatcccctc
17461 ccgcaactgg gcggctttcc gcggctgggc cttcacgcgc cttaagacta aggaaacccc
```

17521 atcactgggc tcgggctacg acccttatta cacctactct ggctctatac cctacctaga
17581 tggaaccttt tacctcaacc acacctttaa gaaggtggcc attacctttg actcttctgt
17641 cagctggcct ggcaatgacc gcctgcctta cccccaaccg agtttgaaat taagcgctca
17701 gttgacgggg agggttacaa cgttgcccag tgtaacatga ccaaagactg gttcctggta
17761 caaatgctag ctaactatta acattggcta ccagggcttc tatatcccag agagctacaa
17821 ggaccgcatg tactccttct ttagaaactt accagcccat gagccgtcag gtggtggatg
17881 atactaaata acaaggacta ccaacaggtg ggcatcctac accaacaaca acaactctgg
17941 atttgttggc taccttgccc ccaccatgcg cgaaggacag gcctaccctg ctaacttccc
18001 ctatccgctt ataggcaaga ccgcagttga cagcattacc caagaaaagt ttactttgcg
18061 atcgcaccct ttggcgcatc ccattctcca gtaactttat gtccatgggc gcactcacag
18121 acctgggcca aaaccttctc tacgccaact ccgcccacgc gctagacatg actttgaggt
18181 ggatcccatg gacgagccca cccttcttta ttgttttgtt tgaagtcttt gacgtggtcc
18241 gtgtgacacc aagccgcacc gcggcgtcat cgaaaccgtg tacctgcgca cgcccttctc
18301 ggccggcaac gccacaacat aaagaagcaa gcaacatcaa caacagctgc cgccatgggc
18361 tccagtgagc aggaactgaa agccattgtc aaagatcttg gttggtgggc cattattttt
18421 gggctaccta tgacaagcgc tttcctaggc tttgtttctc cacacaagct cgcctgcgcc
18481 atagtcaata cggccggtcg cgagactggg ggcgtacact ggatggcctt tgcctggaac
18541 ccgcactcaa aaacatgcta cctctttgag ccctttggct tttctgacca gcgactcaag
18601 caggtttacc agtttgagta cgagtcactc ctgcgccgta gcgccattgc ttccttcccc
18661 cgaccgctgt ataacgctgg aaaagtccac ccaaagcgta caggggccca actcggccgc
18721 ctgtggacta ttctgctgca tgtttctcca cgcctttgcc aacctggccc caaacctccc
18781 atggatcaca accccaccat gaaccttatt accggggtac ccaactccat gctcaacagt
18841 ccccaggtac agcccaccct gcgtcgcaac caggaacagc tctacagctt cctggagcgc
18901 cactcgccct acttccgcag ccacagtgcg cagattagga gcgccacttc tttttgttca
18961 cttgaaaaac aatagtaaaa ataatgtact agagacactt tcaataaagg caaattgctt
19021 ttatttgtta cactctcggg tgattattta cccccaccct tgccgtctgc gccgtttaaa
19081 aatcaaaggg gttctgccgc gcatcgctat gcgccactgg cagggacacg ttgcgatact
19141 ggtgtttagt gctccactta aactcaggca caaccatccg cggcagctcg gtgaagtttt
19201 cactccacag gctgcgcacc atcaccaacg cgtttagcag gtcgggcgcc gatatcttga
19261 agtcgcagtt ggggcctccg ccctgcgcgc gcgagttgcg atacacaggg ttgcagcact
19321 ggaacactat cagcgccggg tggtgcacgc tggccagcac gctcttgtcg gagatcagat
19381 ccgcgtccag gtcctccgcg ttgctcaggg cgaacggagt caactttggt agctgccttc
19441 ccaaaaaggg cgcgtgccca ggctttgagt tgcactcgca ccgtagtggc atcaaaaggt
19501 gaccgtgccc ggtctgggcg ttaggatacа gcgcctgcat aaaagccttg atctgcttaa
19561 aagccacctg agcctttgcg ccttcagaga agaacatgcc gcaagacttg ccggaaaact
19621 gattggccgg acaggccgcg tcgtgcacgc agcaccttgc gtcggtgttg gagatctgca
19681 ccacatttcg gccccaccgg ttcttcacga tcttggcctt gctagactgc tccttcagcg
19741 cgcgctgccc gttttcgctc gtcacatcca tttcaatcac gtgctcctta tttatcataa
19801 tgcttccgtg tagacactta agctcgcctt cgatctcagc gcagcggtgc agccacaacg
19861 cgcagcccgt gggctcgtga tgcttgtagg tcacctctgc aaacgactgc aggtacgcct
19921 gcaggaatcg ccccatcatc gtcacaaagg tcttgttgct ggtgaaggtc agctgcaacc
19981 cgcggtgctc ctcgttcagc caggtcttgc atacggccgc cagagcttcc acttggtcag
20041 gcagtagttt gaagttcgcc tttagatcgt tatccacgtg gtacttgtcc atcagcgcgc
20101 gcgcagcctc catgcccttc ctcccacgca gacacgatcg gcacactcag cgggttcatc
20161 accgtaattt cactttccgc ttcgctgggc tcttcctctt cctcttgcgt ccgcatacca
20221 cgcgccactg ggtcgtcttc attcagccgc cgcactgtgc gcttacctcc tttgccatgc
20281 ttgattagca ccggtgggtt gctgaaaccc accatttgta gcgccacatc ttcttctttc
20341 ttcctcgctg tccacgatta cctctggtga tggcgggcgc tcgggcttgg gaggaagggc
20401 gcttcttttt ctttcttggg cgcaatggcc aaatccgccg ccgaggtcga tggccgcggg
20461 ctgggtgtgc gcggcaccag cgcgtcttgt gatgagtctt cctcgtcctc ggactcgata
20521 cgccgcctca tcctgctttt tgggggcggc ccggggaggg cggcggcgac ggggagcggg
20581 gacgacacgt cctccatggt tgggggagcg tcgcgccgca ccgcgtccgc gctcgggggt
20641 ggtttcgcgc tgctcctctt cccgactggc catttccttc tcctataggc agaaaaagaa
20701 tcatggagtc agtcgagaag aaggacagcc taacccgccc cctcctgagt tcgccaccac
20761 cgcctccacc gatgccgcca acgcgcctac caccttccc cgtcgaggca cccccgcttg
20821 aggaggagga agtgattatc gagcaggacc caggttttgt taagcgaaga cgacgaggac
20881 cgctcagtac caacagagga ataaaaagca aagaaccagg acaacgcaga ggcaaacgag
20941 gaacaagtcg ggcgggggga gcgaaaggca tggcgactac ctagatgtgg gagacgacgt
21001 gctgttgaag catctgcagc gccagtgcgc cattatctgc gacgcgttgc aagagcgcag
21061 cgatgtgccc ctccgccata gcggatgtca gccttgccta cgaacgccac ctattctcac

```
21121 cgcgcgtacc ccccaaaccg ccaagaaaac ggcacatgcg agcccaaccc gccgcctcaa
21181 cttcctaccc cgtatttgcc gtgccagagg tgcttgccac ctatcacatc tttttccaaa
21241 actgcaagat acccctatcc tgccgtgcca accgcagccg agcggacaag cagctggcct
21301 tgcggcaggg cgctgtcata cctgatatcg cctcgctcaa cgaagtgcca aaaatctttg
21361 agggtgcttg gacgcgacga gaagcgcgcg gcaaacgctc tgcaacaagg aaaacaagcg
21421 aaatagaaag tcactctgga gtgttggtgg aactcgaggg tgacaacgcg cgcctagccg
21481 tactaaaacg cagcatcgag gtcacccact ttgcctaccc ggcacttaac ctacccccca
21541 aggtcatgag cacagtcatg agtgagctga tcgtgcgccg tgcgcagccc ctggagaggg
21601 atgcaaattt gcaagaaaca aacaagagga gggcctaccc gcagttggcg acgagcagct
21661 agcgcgctgg cttcaaacgc gcgagcctgc cgacttggag gagcgacgca aactaatgat
21721 ggccgcagtg ctcgttaccg tggagcttga gtgcatgcag cggttctttg ctgacccgga
21781 gatgcagcgc aagctagagg aaacattgca ctacaccttt cgacagggct acgtacgcca
21841 ggcctgcaag atctccaacg tggagctctg caacctggtc tcctaccttg gaattttgca
21901 cgaaaaccgc cttgggcaaa acgtgcttca ttccacgctc aagggcggag gcgcgccgcg
21961 actacgtccg cgactgcgtt tacttatttc tatgctacac ctggcagacg gccatgggcg
22021 tttggcagca gtgcttggag gagtgcaacc tcaaggagct gcagaaactg ctaaagcaaa
22081 acttgaagga cctatggacg gccttcaacg agcgctccgt ggccgcgcac ctggcggaca
22141 tcattttccc cgaacgcctg cttaaaaccc tgcaacaggg tctgccagac ttcaccagtc
22201 aaagcatgtt gcagaacttt aggaacttta tcctagagcg ctcaggaatc ttgcccgcca
22261 cctgctgtgc acttcctagc gactttgtgc ccattaagta ccgcgaatgc cctccgccgc
22321 tttggggcca ctgctacctt ctgcagctag ccaactacct tgcctaccac tctgacataa
22381 tggaagacgt gagcggtgac ggtctactgg agtgtcactg tcgctgcaac ctatgcaccc
22441 cgccaccgct ccctggtttg caattcgcag ctgcttaacg aaagtcaaat tatcggtacc
22501 tttgagctgc agggtccctc cgcctgacga aaagtccgcg gctccggggt tgaaactcac
22561 tccggggctg tggacgtcgg cttaccttcg caaatttgta cctgaggact accacgccca
22621 cgagattagg ttctacgaag accaatcccg cccgcctaat gcggagctta ccgcctgcgt
22681 cattacccag ggccacattc ttggccaatt gcaagccatc aacaaagccc gccaagagtt
22741 tctgctacga aagggacggg gggtttactt ggacccccag tccggcgagg agctcaaccc
22801 aatccccccg cccgccgcag ccctatcagc agcagccgcg ggcccttgcc ttcccaggat
22861 ggcacccaaa aagaagctgc agctgccgcc gccacccacc ggacgaggag gaatactggg
22921 acagtcaggc agaggaggtt ttggacgagg aggaggagga catgatggaa gactgggaga
22981 gcctagacga ggaagcttcc gaggtcgaag aggtgtcaga cgaaacaccg tcaccctcgg
23041 tcgcattccc ctccgccggc cgccccagaa atcggcaacc ggttccagca tggctacaac
23101 ctccgctcct caggcgccgc cggcactgcc cgttcgccga cccaaccgta gatgggacac
23161 cactggaacc agggccggta agtccaagca gccgccgccg ttagcccaag agcaacaaca
23221 gcgccaaggc taccgctcat ggcggcgggc acaagaacgc catagttgct tgcttgcaag
23281 actgtggggg caacatctcc ttccgcccgc cgctttcttc tctaccatca cggcgtggcc
23341 ttcccccgta acatcctgca ttactaccgt catctctaca gcccatactg caccggcggc
23401 cggcggcagc aaagaacagc aacagcagcg gccacacaga agcaaaggcg accggatagc
23461 aagactctga caaagcccaa gaaatccaca gcggcggcag cagcaggagg aggagcgctg
23521 cgtctggcgc ccaacgaacc cgtatcgacc cgcgagctta gaaacaggat tttcccactc
23581 tgtatgtcta tatttcaaca gagcaggggc caagaacaag agctgaaaat aaaaaacaag
23641 gtctctgcga tccctccacc cgcagctgcc tgtatcacaa aagcgaagat cagcttcggc
23701 gcacgctgga agacgcggag gctctcttca gtaaatactg cgcgctgact cttaaggact
23761 agtttcgcgc cctttctcaa atttaagcgc gaaaactacg tcatctccag cggccacacc
23821 cggccgccag ccacctgctt gttgtcagcg ccattatgag caaggaaatt cccaccgccc
23881 tacatgtgga gttaccagcc acaaatggga cttgcggctg gagctgccca agactactca
23941 acccgaataa actacatgag cgcgggaccc cacatgatat cccgggtcaa cggaataacg
24001 cgcccaccga aaccgaattc tcctggaaca ggcggctatt accaccacac ctcgtaataa
24061 ccttaatccc cgtagttggc ccgcctgccc tggtgtacca ggaaagtccc gcctcccacc
24121 actgtggtac ttcccagaga cgcccaggcc gaagttcaga tgactaactc aggggcggca
24181 gcttgcgggc ggctttcgtc acagggtggc ggtcgcccgg gcagggtata actcacctga
24241 caatcagagg gcggaggata ttcagactca acgacgagtc ggtgagctcc tcgcttggtc
24301 tccgtccgga cgggacattt cagatcggcg gcgccggccg cctcttcatt cacgcctcgt
24361 caggcaatcc taactctgca gacctcgtcc tctgagccgc gctctggagg cattggaact
24421 ctgcaattta ttgaggagtt tgtgccatcg gtctacttta acccccttcct cgggacctcc
24481 cggccactat ccggatcaat ttattcctaa ctttgacgcg gtaaaggact cggcggacgg
24541 ctacgactga atgttaagtg gagaggcaga gcaactgcgc ctgaaacacc tggtccactg
24601 tcgccgccac aagtgctttg cccgcgactc cggtgagttt tgcttacttt gaattgcccg
24661 aggatcatat cgagggcccg gcgcacggcg tccggcttac cgcccaggga gagcttgccc
```

Human BAC Sequence and Assembly by 454 Life Sciences' Sequencing by Synthesis Method

Lei Du[1], David A. Wheeler[2], George Weinstock[2], Richard Gibbs[2], Gina Costa[1], John Leamon[1], Jan Berka[1], Srinivasan Maithreyan[1], Gary Sarkis[1], Kenton L. Lohman[1]

[1] 454 Life Sciences, Branford, CT

[2] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

454 LIFE SCIENCES

Abstract

454 Life Sciences has developed proprietary methods for massively parallel DNA sequencing. We have applied this technology to re-sequencing and mapping human BAC clones to their precise chromosomal locations. This preliminary data shows the efficacy of the technology to rapidly sample and characterize subsets of sequence spanning an entire genome or a specific chromosomal location. The novel DNA sequencing method consists of three steps: template preparation, solid phase amplification and solid phase DNA sequencing. Several thousand to several hundreds of thousands of DNA sequencing reactions are performed simultaneously on glass plates containing 300 thousand to 1 million, 75 picoliter volume wells. Average read length of each fragment is consistently greater than 50 bases. The starting point for genome sequencing involves a single-tube template preparation and an absence of a bacterial plasmid cloning step, thus greatly reducing costs and increasing the throughput of our system. In addition, we are completing development of a new software algorithm for de novo whole genome assembly. Sequencing results from human BAC clones will be presented and discussed.

Experimental Procedures

Our novel methodology requires only a single sample preparation per genome, utilizes simultaneous clonal amplification of shotgun fragments in sub-nanoliter microreactors, without the use of time-consuming cloning steps. The product of each microreactor is driven to and captured by a concomitant solid support. The captured DNAs are delivered to wells on the PicoTiterPlate™ and sequenced on 454 Life Sciences sequencing platform. The details of these steps are illustrated in Figure 1.

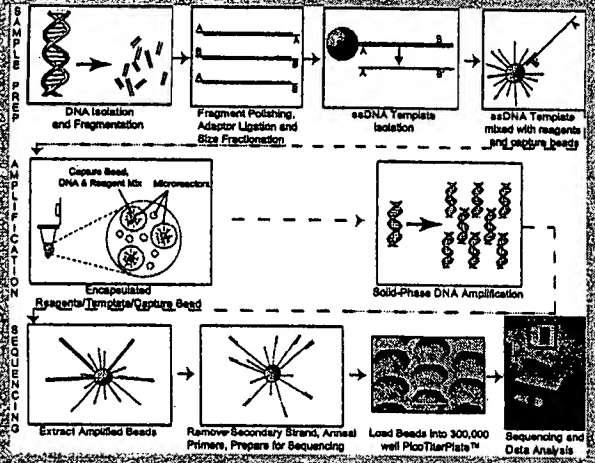


Figure 1: Streamlined template preparation and amplification process.

1) BAC DNA from clone RP11-418C2 was fragmented to sub-kilobase lengths.
2) The fragment ends were polished, 5' and 3' adaptors ligated onto each fragment, and the sample was size fractionated, resulting in products under 500 bases in length.
3) One strand of these double-stranded products was bound to microparticles, and the free strand was eluted as template for the subsequent amplification reaction.
4) Amplification was conducted in a single-reaction preparation, encapsulating the reaction reagent mix, a single DNA capture bead, and template in a 40 to 100 picoliter microreactor.
5) The particular template molecule contained in each individual microreactor was amplified and immobilized on the respective DNA capture bead.
6) The DNA capture beads were extracted and the template DNA was prepared for use on the 454 sequencer.

Data Analysis

The 454 sequencer generates raw traces for each microreactor and produces sequence reads in FASTA format using a proprietary basecaller program. Adaptors and low quality reads are removed and repeats masked before mapping and assembly.

Human Genome Mapping:
Each masked read was mapped against the human genome (NCBI build 33) using BLAT and the mapped reads (>95% identity) are recorded for each chromosome.

BAC Assembly:
Each sequence was mapped against the reference BAC sequence (RP11-418C2) using a proprietary alignment algorithm and the resulting alignment was recorded. For sequences that map to the genome with >90% accuracy, the software generates a list of individual bases found at a given position in the reference genome. The consensus base for each location was computed by averaging all mapped bases. This consensus sequence was then compared with the reference sequence to calculate total accuracy and coverage.

We also mixed 3x oversample of reads (950 sequences) generated from conventional Sanger method with reads generated from the 454 sequencer and assembled with Phrap using default parameters.

Results and Discussion

Human Genome Mapping:
Out of 8561 mapped reads, 7153 are mapping to human chromosome 12 (Fig. 2a). Of these, 7058 reads map to the expected location within chromosome 12 (Fig. 2b). The coordinate boundaries for clone RP11-418C2 in NCBI build 33 are 11,818,492 to 11,986,440, whereas boundaries on the 7058 read stack are 11,816,616-11,986,514. We also mapped these reads to the mouse genome, and located the BAC to the syntenic region on mouse chromosome 6 (data not shown).

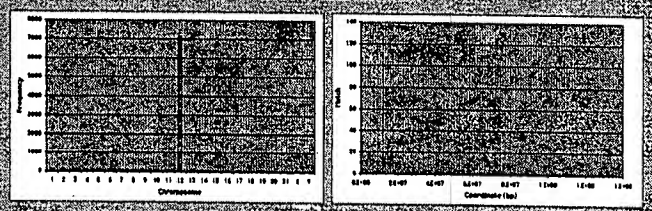


Figure 2: Sequence mapping against human genome and within chromosome 12.

BAC Assembly:
In a separate sequencing run, we generated 57,193 raw reads from this BAC clone. After adaptor removal, repeat masking, and quality trimming, 39900 reads were assembled against the reference sequence (Fig. 3). Genome coverage is 85% and consensus accuracy is 98%. Average read length is 84 bases.

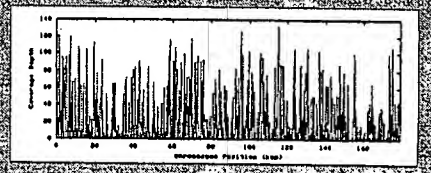


Figure 3: Frequency of assembled reads across BAC sequence length.

Phrap Assembly:
Sanger reads alone generated 25 major contigs (>2 kb) with a 76% mapping efficiency, whereas Sanger and 454 reads combined produced 18 major contigs with a 83% mapping efficiency. 454 reads were able to join and extend Sanger contigs into much larger stretches (Fig. 4).

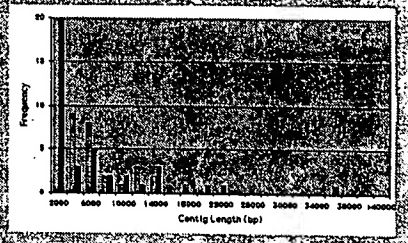


Figure 4: Distribution of Phrap contigs from Sanger only and Sanger+454 reads.

Conclusion

We have demonstrated in this study that 454's Life Sciences' novel sequencing methodology is capable of producing sufficient shotgun sequence coverage of a BAC clone in a single run (done within 1 day). The reads can be used to map its precise location in the genome, as well as assembling into contigs based on a reference sequence. This is a useful tool for whole genome mapping and sequencing.

We also showed that by combining conventional Sanger method with 454 technology, we achieve a better *de novo* assembly outcome for whole genome shotgun sequencing.

We are continuing to develop our quality scoring and trimming algorithm. We have completed phase one of our proprietary fragment assembler, designed to take advantage of the raw trace signals produced by our sequencing-by-synthesis method. This assembler will be available as part of 454's commercial sequencing instrument.

Abstract

454 Life Sciences has developed proprietary methods for massively parallel DNA sequencing. We have applied this technology to re-sequencing and mapping human BAC clones to their precise chromosomal locations. This preliminary data shows the efficacy of the technology to rapidly sample and characterize subsets of sequence spanning an entire genome or a specific chromosomal location. The novel DNA sequencing method consists of three steps: template preparation, solid phase amplification, and solid phase DNA sequencing. Several thousand to several hundreds of thousands of DNA sequencing reactions are performed simultaneously on glass plates containing 300 thousand to 1 million, 75 picoliter volume wells. Average read length of each fragment is consistently greater than 50 bases. The starting point for genome sequencing involves a single template preparation and an absence of a bacterial plasmid cloning step, thus greatly reducing costs and increasing the throughput of our system. In addition, we are completing development of a new software algorithm for *de novo* whole genome assembly. Sequencing results from human BAC clones will be presented and discussed.

Experimental Procedures

Our novel methodology requires only a single sample preparation per genome, utilizes simultaneous clonal amplification of shotgun fragments in sub-nanoliter microreactors, without the use of time-consuming cloning steps. The product of each microreactor is driven to and captured by a concomitant solid support. The captured DNAs are delivered to wells on the PicoTiterPlate™ and sequenced on 454 Life Sciences' sequencing platform. The details of these steps are illustrated in Figure 1.

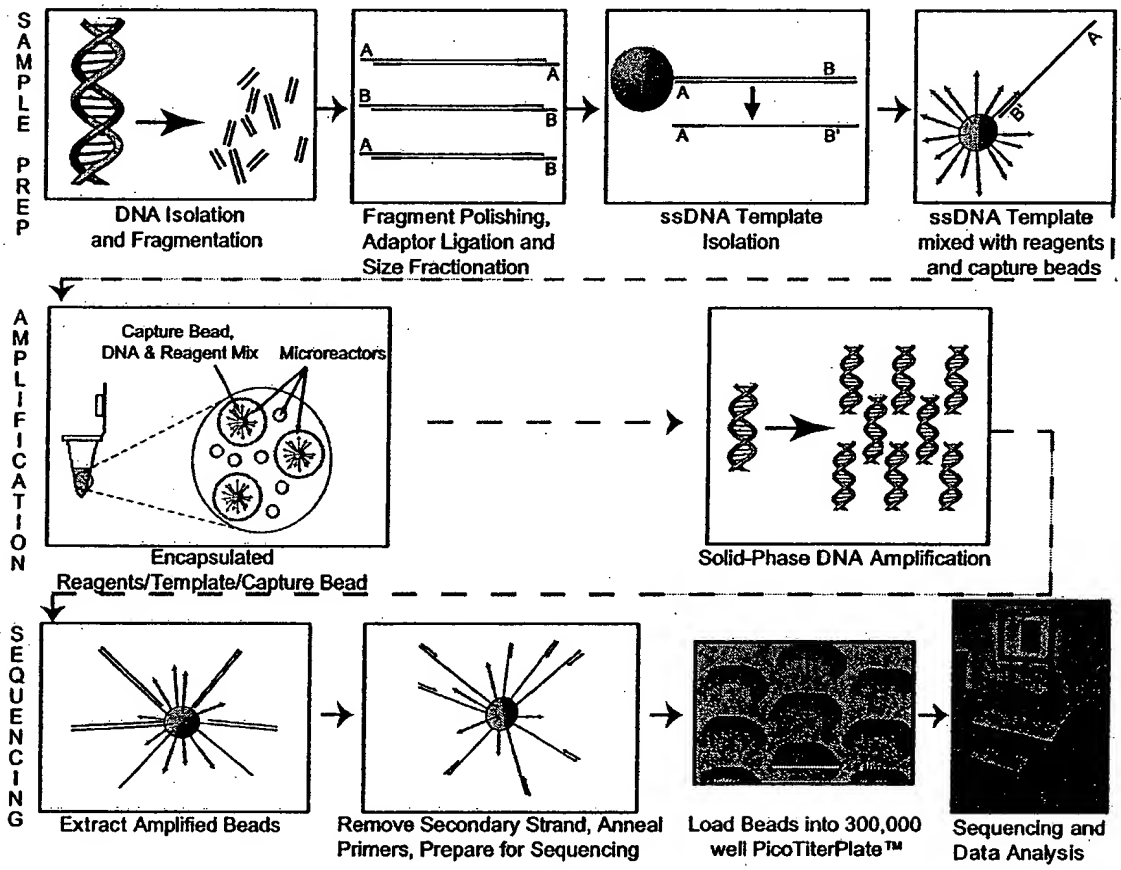


Figure 1. Streamlined template preparation and amplification process

1) BAC DNA from clone RP11-418C2 was fragmented to sub-kilobase lengths.
2) The fragment ends were polished, 5' and 3' adaptors ligated onto each fragment, and the sample was size fractionated, resulting in products under 500 bases in length.
3) One strand of these double-stranded products was bound to microparticles, and the free strand was eluted as template for the subsequent amplification reaction.
4) Amplification was conducted in a single reaction preparation, encapsulating the reaction reagent mix, a single DNA capture bead, and template in a 40 to 100 picoliter microreactor.
5) The particular template molecule contained in each individual microreactor was amplified and immobilized on the respective DNA capture bead.
6) The DNA capture beads were extracted and the template DNA was prepared for use on the 454 sequencer.

Data Analysis

The 454 sequencer generates raw traces for each microreactor, and produces sequence reads in FASTA format using a proprietary basecaller program. Adaptors and low quality reads are removed and repeats masked before mapping and assembly.

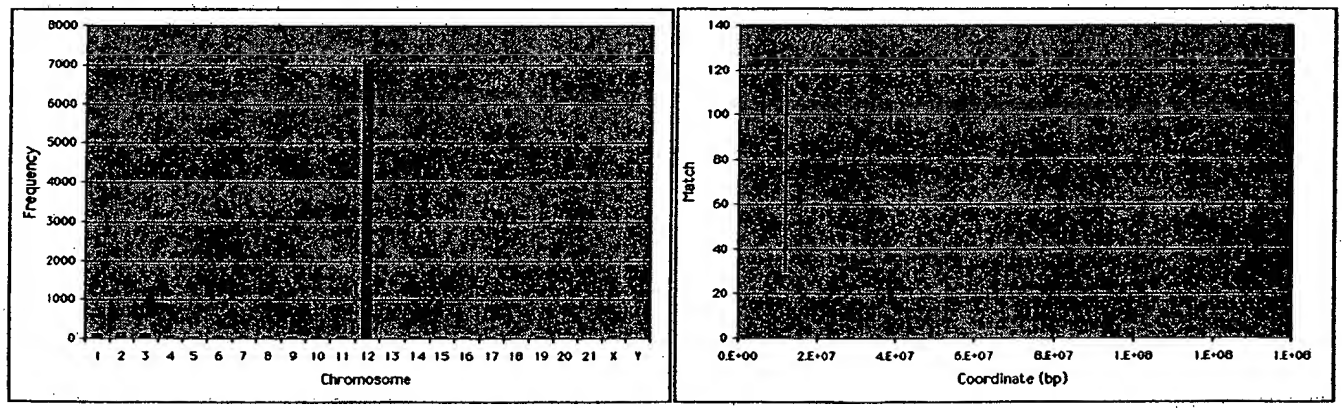
Human Genome Mapping:
Each masked read was mapped against the human genome (NCBI build 33) using BLAT and the mapped reads (>95% identity) are recorded for each chromosome.

BAC Assembly:
Each sequence was mapped against the reference BAC sequence (RP11-418C2) using a proprietary alignment algorithm and the resulting alignment was recorded. For sequences that map to the genome with >90% accuracy, the software generates a list of individual bases found at a given position in the reference genome. The consensus base for each location was computed by averaging all mapped bases. This consensus sequence was then compared with the reference sequence to calculate total accuracy and coverage.

We also mixed 3x oversample of reads (950 sequences) generated from conventional Sanger method with reads generated from the 454 sequencer and assembled with Phrap using default parameters.

Results and Discussion

Human Genome Mapping:
Out of 8561 mapped reads, 7153 are mapping to human chromosome 12 (Fig. 2a). Of these, 7058 reads map to the expected location within chromosome 12 (Fig. 2b). The coordinate boundaries for clone RP11-418C2 in NCBI build 33 are 11,818,492-11,986,440, whereas boundaries on the 7058 read stack are 11,816,616-11,986,511. We also mapped these reads to the mouse genome, and located the BAC to the syntenic region on mouse chromosome 6 (data not shown).



(a) Mapping against Human Genome (b) Mapping to Human Chr12

Figure 2. Sequence mapping against human genome and within chromosome 12

Results and Discussion

BAC Assembly:
In a separate sequencing run, we generated 67193 raw reads from this BAC clone. After adaptor removal, repeat masking and quality trimming, 39900 reads were assembled against the reference sequence (Fig. 3). Genome coverage is 85% and consensus accuracy is 98%. Average read length is 84 bases.

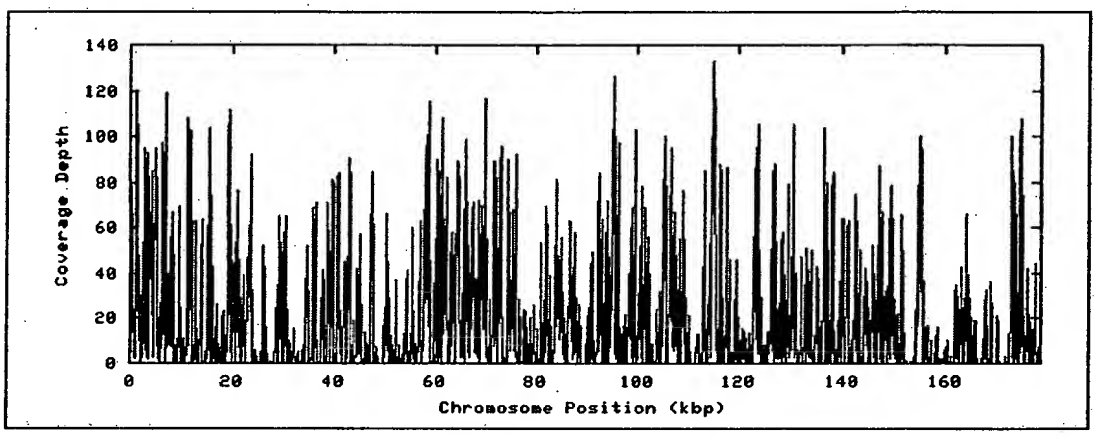


Figure 3. Frequency of assembled reads across BAC sequence length

Phrap Assembly:
Sanger reads alone generated 25 major contigs (>2 kb) with a 76% mapping efficiency, whereas Sanger and 454 reads combined produced 18 major contigs with a 83% mapping efficiency. 454 reads were able to join and extend Sanger contigs into much larger stretches (Fig. 4).

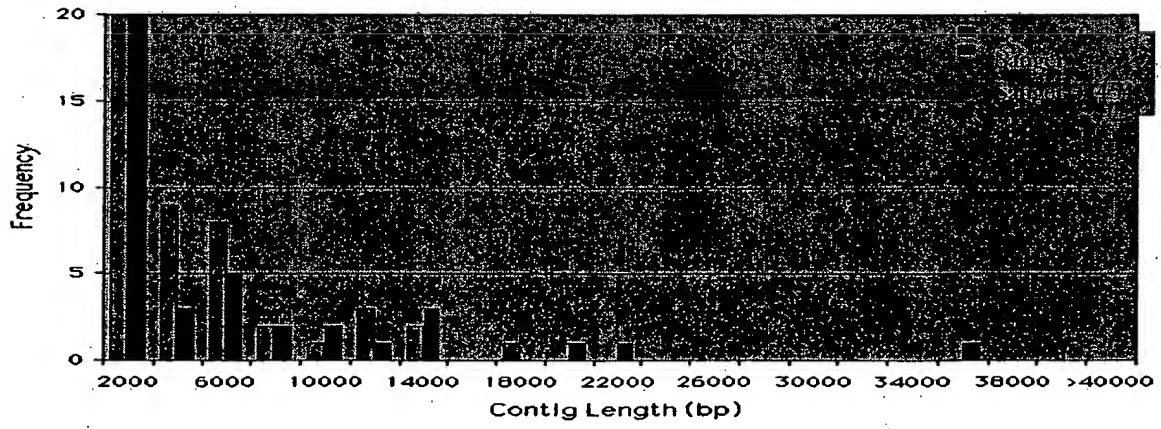


Figure 4. Distribution of Phrap contigs from Sanger only and Sanger+454 reads

Conclusion

We have demonstrated in this study that 454 Life Sciences' novel sequencing methodology is capable of producing sufficient shotgun sequence coverage of a BAC clone in a single run (done within 1 day). The reads can be used to map its precise location in the genome, as well as assembling into contigs based on a reference sequence. This is a useful tool for whole genome mapping and sequencing.

We also showed that by combining conventional Sanger method with 454 technology, we achieve a better *de novo* assembly outcome for whole genome shotgun sequencing.

We are continuing to develop our quality scoring and trimming algorithm. We have completed phase one of our proprietary fragment assembler, designed to take advantage of the raw trace signals produced by our sequencing-by-synthesis method. This assembler will be available as part of 454's commercial sequencing instrument.